

Seeds of consensus

A global agreement on the terms of trade in genetically modified organisms is a victory for common sense. But it also reflects the pragmatic recognition that few would gain from a prolonged trade war over the issue.

Last Saturday's agreement in Montreal of a Biosafety Protocol marks a welcome respite from the hostility that has characterized international relations on the question of the risks and benefits of agricultural biotechnology (see page 473). The detail of the agreement leaves some key points unresolved. But at least it offers a framework for defusing potentially explosive tensions over the international movement of genetically modified (GM) organisms.

Two familiar armies arrived in Montreal to negotiate the agreement. The United States, along with the well-fed representatives of its biotechnology and farming industries, and a small number of food-exporting allies came to argue for minimal restrictions on the import and export of GM organisms. The European Union, a motley crew of environmental activists and most of the developing world came looking for a protocol that would echo their reluctance to accept the safety of such a trade. After last year's failures to agree a biosafety protocol at Cartagena, Colombia, or a format for discussing biotechnology at the World Trade Organization (WTO), the chances of agreement looked slim.

But, perhaps surprisingly, an agreement was struck. The Cartagena Protocol on Biosafety — so named to prevent any confusion with the Montreal Protocol on ozone depletion — sets out new conditions for transfer of live GM organisms between countries, allowing the importing country to make the decision.

The protocol was originally proposed almost a decade ago on the grounds that such transfers were seen as a threat (albeit a small one) to biodiversity. It became tougher to negotiate after environmentalists had successfully argued that it should cover not only live plants and seeds, but also imports of commodity grain, which, they said, could escape accidentally into the environment.

The new protocol sets fairly comprehensive controls on the movement of living organisms, as well as much narrower ones for grain. Shipments of GM grain will have to be labelled as such; but the labelling regime won't come into force until two years after the protocol does, which probably means 2003 or 2004. By then, the labelling issue will probably have been solved by market forces.

In defining the circumstances in which importing countries can decline to accept GM organisms, the protocol draws a fine balance. It rejects the need for "scientific certainty" in order for a party to make such a decision — a move being described by some environmentalists as a triumph for the precautionary principle which, they say, is espoused here for the first time in an international treaty. But the treaty also requires importers to use scientific risk assessment, including cost-benefit analysis, to lay grounds for exclusion.

On the face of it, allowing for labelling — even vague labelling, in three or four years' time — is a major concession by the United States and the agricultural biotechnology industry. This concession was surely made, in part, because of the knowledge that failure to reach any kind of agreement would augur very badly for international trade in GM organisms. But it was balanced by the protocol's own admission that it does not supersede parties' rights under the WTO. The WTO has enforcement mechanisms, where the new protocol has none.

The United States may recognize for now that bringing out the big stick — making a complaint to the WTO — on this issue would be counter-productive, turning Europe's teething troubles over GM food into a blanket rejection, with a trade war thrown in. Constructive engagement in the arcane art of scientific risk assessment is better than that. ■

Electronic submissions now welcome

From this week, *Nature* encourages authors to send all manuscripts in one of a range of digital formats.

Whatever the prophets may have said, the paperless office is nowhere in sight. But in *Nature*'s offices, at least, the sheer quantity of accumulated paper is set to diminish progressively. For we are now able to receive electronic versions of manuscripts of submitted Articles and Letters from all disciplines, and to handle the documents through refereeing, discussion and editing, entirely electronically (we already accept electronic pre-submission enquiries.)

Many specialist journals have offered such a facility for some time. But since *Nature* is a general journal, the range of disciplines covered, with their respective and diverse established formats, has impeded its ability to open the electronic hatch to all comers, even though pilot experiments have been working successfully for several months.

No one should conclude that those submitting on paper will be discriminated against in any way. Nor will potential referees be excluded because they cannot handle a submission in electronic form. The publication process may prove slower on average for paper submissions, but only for the inescapable practical reasons that the

Internet is faster, more readily usable and more accessible by authors and referees than fax transmission, let alone the post.

Details of our online services for authors are gathered together in one virtual location: <http://www.nature.com/submit> (for a quick tour of what we can offer, see overleaf). Alternatively, simply e-mail nature@nature.com for instructions. Online submission is based on three stages: information via the *Nature* website; transmission via WAM!NET, a dedicated high-speed network providing reliable and secure file transfer using the Internet; and conversion to Adobe Portable Document Format (PDF) for assessment and refereeing.

Over the coming weeks, we will extend this facility to encompass Brief Communications. Over the months and years ahead, we expect burgeoning bandwidths to play their part in encouraging an increasing amount of electronic submission and refereeing. We expect, too, to be able to handle a greater variety of digital formats, both known and as yet unknown. Meanwhile, authors looking for the greatest speed of response from *Nature* are encouraged to use the new system. ■





Webbed feat
Researchers launch
online amphibian
conservation project
p471



Biosafety
Surprise agreement
in Montreal on trade
rules for GM food
p473



Germany's past
Max Planck Society
comes to terms with
its Nazis
p474



Roving robot
Nomad hunts for
meteorites in the
Antarctic wastes
p477

Livermore plans radical rescue for 'mismanaged' laser facility

Livermore, California

Bruce Tarter, the director of the Lawrence Livermore National Laboratory in California, will this week announce a major reorganization of the laboratory aimed at countering criticism that it has mismanaged the construction of the National Ignition Facility (NIF), a huge laser facility being built there.

When major problems with the project first emerged last September (see *Nature* 401, 101; 1999), cost overruns on the NIF's original construction budget of \$1.2 billion were unofficially estimated at \$300 million. Sources close to the project now put the overrun at \$500 million.

Tarter will split Livermore's laser division, which accounts for more than a quarter of the laboratory's annual \$1.2 billion budget, and appoint an associate director directly responsible for the project's completion. He has also pledged to provide more direct leadership for the NIF. "I expect to be a public spokesman. This project is a major priority for the institution — perhaps my top priority — and I expect others at the laboratory to reflect that."

The Secretary of Energy's Advisory Board (SEAB) is examining the project but has not publicly estimated its eventual cost. However, Ed Moses, NIF project manager at Livermore, says SEAB expects overruns of around \$500 million. Hermann Grunder, director of the Thomas Jefferson Accelerator Facility in Virginia, says the original \$1.2 billion budget was unrealistic. Grunder, a member of the NIF Council, an outside advisory body also examining the project, says that the NIF "always was a \$1.7 billion project".

External scientists who are examining the project's progress believe that its final construction cost could be as high as \$2 billion. They believe, however, that the NIF will be completed to specification, and that technical problems highlighted in a recent SEAB report will be overcome.

Government officials in Washington believe that, at this level, the NIF will run into major difficulties, delaying its 2003 completion date by several years. In President



Tarter: will be NIF's 'public spokesman'.

Clinton's budget proposal for 2001, to be released next Monday, the project's budget allocation is expected to fall sharply from this year's, in line with the original construction schedule.

Rather than obtaining more money from the administration to cover overruns, Livermore is preparing a new cost estimate, and will look to Congress for extra money to prevent the project from falling too far behind schedule.

However, the nuclear weapons budget is already tight, with lay-offs likely at many weapons research and production facilities, and some observers doubt that extra money will be found for the NIF. Two House committees have asked the General Accounting Office to investigate the project, and may hold hearings on its progress in early March.

Under its new project plan, Livermore will appoint an outside engineering contractor to oversee the assembly of the NIF lasers, a task originally allocated to the laboratory itself. Tarter says the laboratory will only be able to estimate the project's full cost when it

has taken bids from contractors for the assembly job.

The problem of assembling the NIF's 192 giant lasers in the necessary clean conditions is compounded by the fact that they are to be packed tightly inside the NIF building, like well-aligned matches in a full box. "The assembly is something we did not do a good job of forecasting," concedes Tarter.

According to NIF project director Moses, the complexity of the task was overlooked during the design, which sought to keep costs down by making the building as compact as possible, ignoring the difficulties of assembly. Construction of the NIF is proceeding as planned, with a massive concrete structure ready to accommodate the 24 bundles of eight football-field-length lasers. These lasers will fire 1.8 megajoules of energy into a tiny deuterium-tritium target when the facility is in operation (see below).

The NIF's target chamber is also now in place, with land set aside next to it for a second chamber which may be built by Britain's Atomic Weapons Establishment.

Published reports on the NIF and interviews with officials involved in the project reveal a catalogue of errors that have taken the project over its budget. These include:

Nuclear fusion without weapons testing

The importance of the US National Ignition Facility (NIF — see above) is endorsed by nuclear weapons scientists who want to investigate the conditions inside a nuclear explosion without testing, and by advocates of fusion energy. It will be the first device to ignite a controlled fusion reaction.

By focusing an enormous 1.8 megajoules of energy on a pellet of deuterium-tritium fuel a couple of millimetres in diameter, in a pulse lasting only three nanoseconds, the NIF should make the pellet implode, causing

its centre to 'ignite' in a brief, self-sustaining fusion reaction.

Fusion in the pellet could be induced either directly, by firing lasers at it from all sides, or indirectly, by placing it in a cylindrical target, or 'hohlraum', about one centimetre long, and using the laser pulse to induce X-rays from the hohlraum which would compress the pellet.

Even short of ignition, laser experiments such as the NIF cause some fusion in their targets. Nuclear weapons scientists can therefore study the behaviour of



The target, or 'hohlraum'.

materials in extreme conditions, and refine computer models of nuclear explosions without nuclear testing.

C.M.

- External reviewers knew that the project couldn't be built for \$1.2 billion, but declined to say so in public. According to Grunder, who served on a National Academy of Sciences panel that reviewed the NIF proposal in 1997, the panel believed that the project would cost \$1.7 billion to complete, but didn't put the figure in its published report.
- In an effort to control costs, the NIF's 192 lasers are packed tightly together. Previous laser designs, such as Livermore's Nova facility, have been spread out on the ground, allowing technicians easy access during assembly. Clean assembly of NIF's lasers

will therefore be very difficult, although independent assessors say it can be done. "We did it so that the NIF could go in a realistically sized building," says Tarter. "We didn't appreciate the fact that if you packed the lasers closely together, lab technicians couldn't go in" and work on them during assembly.

- Research efforts, such as the Beamlet laser and a laser amplifier called Amplab, which were to have run in parallel with NIF construction and to answer questions critical to its success, were discarded early because of budgetary constraints. Tarter says Beamlet — a full-sized prototype of a NIF laser — should have been kept running as a test bed for NIF technologies, instead of being shut down once it was established that it worked.

- Other research efforts required to complete the project successfully, including investigation of the target 'hohlraums' for the facility and several issues related to the durability of its optical components, are short of resources. "The programme is in deep trouble in terms of the money it needs in very many areas," says one well-informed administration official in Washington.

On the plus side, the optical components identified by SEAB as falling furthest short of the necessary technical specification will be installed near the target chamber, and won't have to be slotted in until late in the construction schedule, giving researchers more time to improve their specification.

As part of a nuclear weapons programme unaccustomed to public scrutiny during the



Moses: upbeat.

Cold War, the NIF received little outside supervision — or even supervision by Tarter — until last summer, by which time the project was in deep trouble. Government officials and scientists from outside the weapons labs believe the NIF lacked

the accountability now expected of publicly funded projects. As Grunder puts it: "Times had changed, but the people [in the weapons programme] hadn't".

A potential candidate for associate director after the reorganization is Moses. He currently reports to George Miller, Livermore's associate director for national security, who took on supervision of the NIF when the director of the project, Michael Campbell, resigned last September.

Strolling along the smooth concrete floor of the facility, which he describes as "the largest optical table in the world", Moses expresses confidence that a new contractor will solve the assembly problem. "We'll be working with people who built pharmaceutical and semiconductor plants," he says.

Moses, an electrical engineer with a background in project management, is strongly upbeat about the NIF construction, but won't speculate on how much it will eventually cost. "We're working through it as accurately as we can," he says. **Colin Macilwain**



Costly: but NIF construction is proceeding steadily.

Japan may allow human embryo stem-cell research

Tokyo

Research on human embryonic stem cells has come a step closer to being legalized in Japan, as the country's science policy-makers prepare a proposal to the government endorsing such research.

A subcommittee of the Council for Science and Technology (CST), Japan's principal science policy-making body, is expected to submit its proposal to the CST by the end of the month. The proposal should receive approval by the end of March, then be submitted to the government.

The Science and Technology Agency, which manages the CST, will issue guidelines on cloning research and human genetic engineering at the same time.

The use of embryo cells for research is expected to be allowed subject to several conditions. First, that only cells from fertilized eggs less than 14 days old are used. Second, that no money is paid to donors for cells. Third, that research is limited to cells that were not used in fertility treatment.

Under a two-tiered approval system, an internal review board of the institution

where the research is to be carried out must endorse the research group's plan before sending it to the government for approval.

A major ethical split between obstetricians and prospective stem-cell researchers on the committee was resolved when it was decided that information about embryo donors would not accompany donated cells. Debates continue on the commercialization and patenting of research results.

The issue has stirred up little public interest, especially compared with last year's intense debates over organ transplants from brain-dead donors. But some researchers have voiced concern over the prospect of starting research before adequate regulatory laws or organizations are in place. One member of the stem-cell subcommittee argues that the CST is too weak. He wants to see an information protection law set up, with an authoritative body to oversee genetic and embryo research.

In the absence of these, he says, "research ethics in human embryology and genetics might be handed down to

individual researchers or institutes, in which case unethical research protocols might be approved. As a consequence, public trust could be lost." This would undermine future research initiatives.

A Ministry of Health and Welfare working group will issue its own guidelines on the use of human genetic information by March (see *Nature* 402, 8; 1999), and the Ministry of Education, Sports and Culture (Monbusho) plans to set up an ethics review committee on human genome research within its managerial board in April.

Lack of a clear central policy, coupled with separate ministerial attempts to assert their own policy-making authority, has long characterized Japanese government activities. Resultant uncertainty in the private sector could severely hamper Japan's plans to lure venture businesses into biotechnology and create 1,000 start-up companies by 2010 (see *Nature* 397, 554; 1999).

"Unlike in the United States, Japanese businesses are not eager to take the initiative when legal and regulatory matters remain unsettled," one source warns. **David Cyrano**

Change of government bodes well for NZ science funding

Wellington

Researchers in New Zealand are optimistic that the country's new Labour-led coalition government will give science more support than it has received for a decade.

The previous National Party government, led for the past two years by prime minister Jenny Shipley, had argued that competition would deliver more science without increasing investment. But scientists in universities became demoralized at the low level of success in applications for research grants.

Under the National Party, the Department of Scientific and Industrial Research was broken up into nine government-owned companies called Crown Research Institutes (CRIs), which compete for funds (see *Nature* 391, 426; 1998).

The new prime minister, Helen Clark, is giving priority to repairing the university sector. But the Labour Party is politically moderate and ministers are wary of major reorganizations.

Clark has given responsibility for higher education, as well as the sensitive portfolio of social security, to Steve Maharey, a former sociology lecturer. Maharey plans to set up a Tertiary Education Advisory Commission this month, which he hopes will provide its first report and recommendations in time to influence the Budget in May and will then keep rolling with frequent reports, rather than putting the system on hold with a comprehensive review.

Pete Hodgson, the minister for research, science and technology, is responsible for 'contestable' grants and the CRIs. Following



Hood: 'benchmarks must be global'.

criticism of the lack of contact with his predecessor, Maurice Williamson, Hodgson is setting up a Science and Innovation Council of external advisers. The council will report to the prime minister. But, like the tertiary commission, its terms of reference and membership have yet to be settled. Hodgson, one of Clark's key strategists, stresses that the research, science and technology effort will be part and parcel of economic and social policy, whereas it used to be marginal. A boost to the support for research and development will be declared in the Budget, sharing in NZ\$1.2 billion (US\$590 million) already identified for new spending across all portfolios.

Some scientists are concerned that the government could have too many sources of advice — from the ministry, the grant-



Maharey: wants continuous reform.

giving Foundation for Research, Science and Technology, the Royal Society of New Zealand, and the new council. Maharey says that coordination will be achieved by a high-level Cabinet committee, bringing him together with Hodgson, Michael Cullen, the powerful finance minister, and Jim Anderton, the minister for industry and, as leader of the left-leaning Alliance Party in the coalition, deputy prime minister.

John Hood, the vice-chancellor of Auckland University — the nation's largest — stresses that the advisory bodies need to be resourced with secretariats and investigative capacity independently of the relevant ministries. He says the previous system failed to make valid international comparisons, and wants reference groups of experts to be appointed from outside New Zealand.

Maharey agrees with this prescription, and will invite universities to nominate their own strengths.

Peter Gluckman, Auckland University's dean of medicine, who chairs the Independent Biotechnology Advisory Council appointed by the National Party, is concerned that "the urgent need for deep analysis of policy and perversities in the system may not be accepted" by the Labour Party. "We cannot tolerate only fine-tuning — the CRIs are in urgent need of review."

The results from ten years of CRIs are mixed. Some, notably Industrial Research Limited, can claim success in generating revenue and spinning-off high-tech companies while maintaining research, for example in high-temperature superconductivity.

This sector generates profits. But the horticulture CRI dismissed some staff just before Christmas, and some CRIs mainly provide technical services to niche industries. Links with universities are expanding but remain patchy. Labour's policy encourages the amalgamation of CRIs.

Andrew West, president of the Association of CRIs and chief executive of Geological and Nuclear Sciences Limited, is calling for "a bold, substantial and sustained increase of public investment in R&D".

Neville Blampied, a psychologist at Canterbury University and the new president of the Association of University Staff, welcomes the abandonment of the "market-driven model". But he is concerned that university heads may not agree on priorities. **Peter Pockley**

WWW project aims to address worldwide decline in amphibians

San Diego

A website designed to describe all amphibian species was launched last week as part of the efforts of an international task force to promote research into the decline in amphibian populations.

The website (<http://www.amphibiaweb.org>) is the latest project by members of the Declining Amphibian Populations Task Force, a worldwide organization of about a hundred working groups (each containing 1–20 scientists) studying the amphibians' decline.

Over the next year, the task force intends to issue a CD database that is planned to include all available data on declining amphibian species.

Frog populations have declined over the past decade, particularly in countries such as Costa Rica, Panama and Australia. As evidence of amphibian decline has mounted, fears have grown that the decline is caused by one or more global factors, including increased ionizing radiation from ozone depletion, chemical contamination, pathogens or unknown stress factors.

Amphibians are important for ecological and biodiversity studies, giving information on the ecological impact of global change, sometimes with implications for humans. Formed in 1992, the task force is a grass-roots response to declining amphibian populations. It has one employee based at the Open University in England, and can be reached through its website (http://www.open.ac.uk/OU/Academic/Biology/J_Baker/JBtxt.htm).

"The task force is a bottom-up organization that started with idealism," says charter member David Wake, professor of integrative biology at the University of California, Berkeley, and a curator of the university's Museum of



▶ Vertebrate Zoology, whose collection forms the core of the group's website.

Wake, the driving force behind the website, said the long-term goal is to develop a page for each of the nearly 5,000 species of frogs, toads, newts, salamanders and caecilians (burrowing, wormlike amphibians).

"The release of amphibiaweb.org is a plea for volunteers," says Wake. "If scientists see a species missing, help us start a page."

After six years as chairman of the task force, Ron Heyer, curator of amphibians and reptiles at the National Museum of Natural History in Washington, is stepping down this spring. His successor is expected to be James Hanken, professor of organismic and evolutionary biology at Harvard University and herpetology curator at the university's Museum of Comparative Zoology.

Balloting began last week on Hanken's appointment. The handover of responsibilities is expected to take place in June, when scientists meet in La Paz, Mexico, for the annual conference of the American Society of Ichthyologists and Herpetologists and other organizations.

Reflecting on his tenure, Heyer says that the most important point was that "a consensus was reached that in fact the amphibian population decline and disappearance phenomenon is real."

There was debate about whether the decline was a cyclical downturn, he says, but studies have convinced amphibian biologists that "something catastrophic was happening".

A reflection of the growing interest in amphibian studies is a \$3 million National Science Foundation grant to 21 investigators last autumn. Headed by James Collins, chair of biology at Arizona State University, the three-year international project will examine the role of host-pathogen biology in the decline of amphibians. Collins says that biologists "not accustomed to big science" projects will be involved in a study involving fields from molecular immunology to global climate change.

"If they answer the research questions they are asking," says Heyer, "they will give us all the information we need to determine where disease fits in to the amphibian picture."

There will be a session on amphibian decline at this month's annual meeting of the American Association for the Advancement of Science in Washington.

One study at the meeting will report the discovery of chytrid fungus in salamanders in Arizona. The chytrid fungus is known to kill frogs, and may be one cause of the worldwide decline in frog populations.

Rex Dalton

Anderson steps down from Wellcome over Oxford row

London

The controversy surrounding Oxford zoology professor Roy Anderson deepened last week with the news that he has stepped down temporarily from his responsibilities as governor of the Wellcome Trust. This follows Anderson's recent suspension on full pay from his Oxford post — and his directorship of the Wellcome Centre for the Epidemiology of Infectious Diseases — pending disciplinary hearings.

It has now emerged that three formal complaints have been made against Anderson concerning his alleged attempts to influence a decision by a zoology department appointments committee, which he was chairing, for a readership in epidemiology (see *Nature* 403, 353; 2000).

The complaint by the successful candidate, Sunetra Gupta, came to prominence when remarks made by Anderson became public and she filed a complaint through a solicitor late last year. Angela McLean of the Biotechnology and Biological Sciences Research Council's Institute for Animal Health has also filed a complaint through a solicitor. The third complaint was made by a senior researcher in the department.

The complaints form part of inquiries into Anderson's behaviour both outside and during an appointments committee in September last year. If convened, Oxford's Visitorial

Board — the five dons who will lead the disciplinary hearing — would also consider written evidence from members of the committee.

The Wellcome Trust says Anderson offered to step down temporarily from his £50,000 a year job as trustee — and from his responsibilities on the three Wellcome Trust advisory panels that he works on — because of formal proceedings against him. Mike Dexter, director of the trust, says he does not know the number of complaints against Anderson or their basis. "It was his wish not to take part in any decisions [at the trust], which I think is honourable," says Dexter.

Dexter says Anderson's action is entirely related to the formal complaints, and that the trust is not pursuing any kind of investigation of its own. He refuses to speculate on whether such an investigation might follow, although he does not rule it out.

"It's an Oxford matter at the present," says Dexter. "It would be most unfair and unfortunate if we were to intervene at any stage. Roy's a brilliant scientist, a wonderful servant of the trust, and we simply allow the process to continue."

The appointments committee voted six out of eight in favour of Gupta. Her candidacy was supported by Sir Robert May, another professor in the department and the government's chief scientific adviser and head of the Office of Science and Technology. **Natasha Loder**

Nossal named Australian of the Year

Sydney

Sir Gustav Nossal, a leading immunologist and former president of the Australian Academy of Science, has been appointed Australian of the Year, an honorary post that nevertheless provides a platform for advocacy.

On the announcement of his appointment by the prime minister, John Howard, Nossal declared three issues he would pursue: increased public investment in universities and research (especially in medicine); improvements in the health of Australian Aborigines; and reconciliation between Aborigines and other Australians on social issues, particularly land rights.

Aborigines suffer much higher disease and mortality than most Australians. The Aboriginal community is demanding an apology for the damage to their society since European settlement began in 1788.

Nossal's appointment enhances his influence as deputy chairman of the Council for



Nossal: will promote health of Aborigines.

Aboriginal Reconciliation, charged with producing a declaration and initiating action acceptable to both sides. Howard has been attacked for being slow to respond to Aboriginal protests. Nossal expects Aborigines to disrupt the Olympic Games in Sydney.

Peter Doherty, another immunologist and former Australian of the Year, who won the award after sharing the Nobel Prize for Physiology or Medicine in 1996, used his term to advocate increased funding for medical research. The government approved a major increase last year (see *Nature* 399, 94; 1999). **Peter Pockley**

IBM joins genomics mapping consortium

Paris

The ambitions of IBM, the world's largest information technology company, to become a major player in genomics took a step forward last week. The company joined a consortium set up to produce a map of human genetic markers known as single-nucleotide polymorphisms, or SNPs.

This US\$50-million joint effort by drug companies and academic centres, begun last year, aims to generate a map and place it in the public domain by 2001. IBM is the first core information technology company to join, and will pay a subscription of \$3 million.

"IBM will add sophistication," says Arthur Holden, chairman and chief executive officer of the SNP Consortium. "Bioinformatics has been dominated by biologists with an interest in computing, and genome software is very basic. We really need to harvest more of the mainstream information technology capacities."

Other members include Britain's Wellcome Trust, AstraZeneca PLC, Aventis Pharma, Bayer AG, Bristol-Myers Squibb Company, Hoffman-LaRoche, Glaxo Wellcome PLC, Novartis, Pfizer Inc., Searle, Smith-Kline Beecham PLC and Motorola Inc.

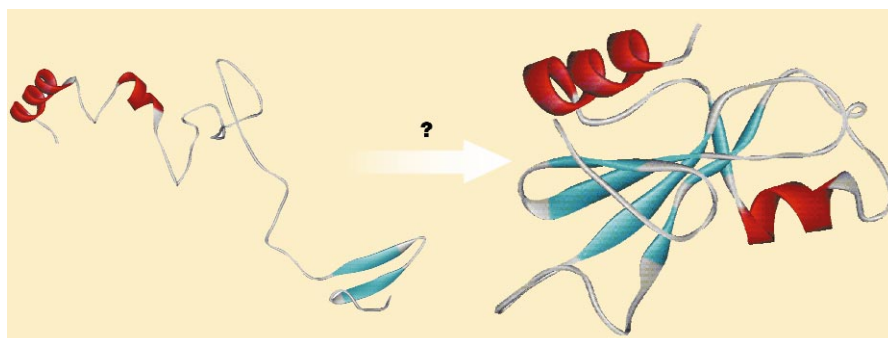
SNPs are identified and analysed in Britain at the Sanger Centre near Cambridge and in the United States at the Whitehead Institute for Biomedical Research, Washington University School of Medicine, Stanford Human Genome Center, and Cold Spring Harbor Laboratory.

The consortium (<http://snp.cshl.org>) aims to identify 300,000 SNPs by April 2001, and has mapped 150,000 of them to their positions on the genome. SNPs occur about once in every 1,000 bases of the 3 billion bases in the human genome. They are key to developing genetic medicine, allowing assessment of individuals' predisposition to diseases, and tailoring therapies.

SNPs will help track down the location of genes in disease, when whole-genome scans of populations susceptible to a disease are compared with those of others that are not. The strategy is to take DNA from 24 ethnically diverse individuals, create small representative libraries across the genome, sequence them and compare the overlapping traces.

It will yield "a broad evenly spaced map across the genome representative of diversity," says Holden. He adds that the built-in diversity will yield large numbers of novel SNPs — one problem is that many of the common SNPs have no role in disease.

To allow time for mapping, the consortium plans quarterly releases into the public domain on the dbSNP database (<http://www.ncbi.nlm.nih.gov/SNP>). The second of these brings the number of mapped SNPs to 7,365.



Protein problem: IBM says that its computing power will help researchers work out how unfolded proteins (such as Barnase, left) become the folded form (right).

The research centres have now scaled up to maximum output, says Holden, and 42,000 SNPs have been identified. This acceleration is "fantastic," says Genghis Lloyd-Harris, director of biotechnology and pharmaceutical equity research at Credit Suisse First Boston in London. It will allay concerns that the project was making slower progress than expected in the face of private rivals such as the US company Celera Genomics and France's Genset, he says.

Dan McCurdy, IBM's vice-president for life sciences, says the decision marks IBM's commitment to a public-domain effort. He adds that, if SNPs are to be broadly applicable to diagnosis and treatment, massive data-handling capacities will be needed to screen an explosion in genetic information.

"Hundreds of genomes are going to be sequenced, and today's tools are relatively crude," says Jeff Augen, IBM's director of life science solutions development. Genomics needs more computer science input and a huge increase in computer power, he says.

In return, IBM will "gain a better understanding of the genome industry by sitting on the board and technical committees of the consortium," he says. It is also interested in

using SNPs in its own research to see how polymorphisms affect protein folding.

The move follows IBM's launch in December of a \$100 million research programme — built around the challenge of modelling protein folding — to build a petaflop supercomputer within five years (see *Nature* 402, 705; 1999). This would carry out more than one quadrillion floating point operations ('flops') per second: some two million times more than the best desktop machines. IBM has also created a life sciences division.

IBM also recently announced \$1 million grants to pilot centres within the US National Institutes of Health's Protein Structure Initiative, a structural genomics programme established last June. The grants are for IBM RS/6000 SP supercomputers and other technologies, while scientists at the centres would have access to software and other resources at its Deep Computing Institute.

It is too soon to say what IBM's licensing policy for software will be, says McCurdy, but he expects it will make some tools freely available to academics. IBM recently made freely available the source code to its Visualization Data Explorer (<http://www.research.ibm.com/dci/software.html>). **Declan Butler**

Rules agreed over GM food exports

Washington

Negotiators from 130 countries last week agreed on a Biosafety Protocol that will require exporters to identify genetically modified (GM) organisms, and allow importing countries to judge whether they pose environmental or health risks.

The surprise agreement bridges deep divisions between the United States and Europe on whether GM and non-GM food should be treated differently for trade purposes. It was reached early in the morning of 29 January, after four days of negotiation in Montreal (see *Nature* 403, 233; 2000).

The agreement states that bulk

shipments of GM foods will be labelled as "containing genetically modified organisms", and that a computer database maintained by the exporter will provide importing countries with information about their contents.

Importers can block shipments, even without "scientific certainty" that a commodity poses a risk. Environmentalists hailed this as a historic breakthrough — the first time that the so-called precautionary principle has been incorporated in an international agreement.

But the agreement — to be called the Cartagena Protocol, after the Colombian

But the agreement — to be called the Cartagena Protocol, after the Colombian city where an earlier round of negotiations ended inconclusively last year — does not claim precedence over the rules of the World Trade Organization (WTO), and exporters who believe they are being treated unfairly will still have recourse to the WTO.

The Biosafety Protocol is the first treaty based on the Convention on Biological Diversity, established at the Rio Earth Summit in 1992. The United States — the largest exporter of GM food — has not ratified the convention, but it played a leading role in the Montreal negotiations and says that it will abide by the protocol.

Although the protocol was conceived chiefly to control the introduction of living organisms into foreign ecosystems, discussions soon became embroiled in arguments about bulk movements of grain. In Cartagena, the United States and some of its food-exporting allies sought to exclude grain shipments from the protocol, arguing that they posed no environmental or health risk, and that a protocol that included commodities would be used to bar imports of US grain.

But observers say that US and European negotiators at Montreal were under pressure to obtain an agreement after the failure of last November's WTO talks in Seattle. Agricultural biotechnology companies, who had encouraged the US to oppose a protocol covering commodities, were prepared to accept the relatively mild controls on commodity shipments: these will not come into force until two years after the protocol is ratified by 50 countries.

Colin Macilwain



Snatching victory? Environmentalists are pleased with the Montreal agreement.

German science begins to cure its historical amnesia

Berlin

Documents relating to the Nazi era from the files of Adolf Butenandt, director of the Kaiser Wilhelm Institute for Biochemistry in Berlin during that time, and later head of the Max Planck Society (MPS), will this week be opened for the first time for examination by historians of science.

The decision to open up the archives reflects a new determination by top MPS officials to come to terms with the wartime activities of scientists who worked for its predecessor, the Kaiser Wilhelm Society (KWS). Butenandt, who won the Nobel prize for chemistry in 1939 for his work on the isolation of sex hormones and went on to develop the one gene—one enzyme hypothesis, is widely revered as a hero of postwar German science.

Although the KWS lost many of its scientific elite during the Third Reich, when Jews and other 'undesirables' were expelled, the MPS has only recently started to address the role of those remaining scientists in supporting Nazi policy. Many continued their careers in the MPS and in universities after the war, and effectively blocked enquiries into their activities and those of their colleagues during the Third Reich. Investigations became taboo.

The taboo is now ending. The MPS has contracted a group of researchers to examine the role of KWS scientists in developing and supporting Nazi policies, and Butenandt's files, which were to have remained closed in the MPS archives for 30 years after his death in 1995, are being made available to them.

But tensions remain. Last week, Robert Proctor, a historian of science from Pennsylvania State University who is working as a guest scientist with the German historians, was refused access to the archives on the grounds that the files had not been properly indexed. However, after intervention from senior MPS officials, Proctor has been assured of access this week.

The MPS's actions contrast with its defensive stance during the 1980s when concerns were raised that, as head of the Berlin institute during the Second World War, Butenandt knew about — and therefore passively colluded with — research on human subjects that members of the KWS were carrying out.

Proctor, whose books include the recently published *The Nazi War on Cancer* (Princeton University Press, 1999), says that historical reflection can only help to normalize perspectives, and he applauds the action taken by the MPS. "The legacy of Nazism influences the negative attitude

many Germans show towards genetics today, and is probably responsible for the absence of a prominent science-orientated intellectual elite: where are the Goulds, Hawkings and Dawkins of Germany?" he asks.

Turning a blind eye

One outspoken critic of the veil of silence that he claims the postwar scientific community drew over the Nazi period is Benno Müller-Hill, a professor of genetics at the University of Cologne. He argues that "it is hard to imagine that a talented science administrator like Butenandt" could not have known, for example, that his colleague Günther Hillmann had analysed blood samples at his Institute for Biochemistry for an experiment at the Auschwitz concentration camp.

Müller-Hill had found evidence that Otmar Freiherr von Verschuer, director of the Kaiser Wilhelm Institute for Anthropology, Human Genetics and Eugenics in Berlin after 1942, worked with Josef Mengele on experiments to identify the hereditary aspects of resistance to infectious diseases.

The experiments involved infecting twins — mostly Jewish — with typhus or tuberculosis and observing the resulting pathologies. Blood samples were sent for analysis to Hillmann, who studied the 'defence enzymes' now known not to exist, but thought at the time to confer resistance to disease (see *Nature* 393, 109–111; 1998).

Müller-Hill was threatened with lawsuits by Butenandt and some of his influential former pupils before he published details of the Hillmann connection in his 1984 book *Tödliche Wissenschaft* (Murderous Science). But the lawsuits did not materialize. And Hubert Markl, the current MPS president, has acted to ensure that the society examines its past systematically, partly in response to pressure from the society's new generation of science historians.

In particular, the MPS is supporting a five-year research programme, 'The Kaiser Wilhelm Society during National Socialism', being carried out by seven external historians of science, and by visiting scientists. The group, headed by Doris Kaufmann, a professor of history at the Technical University of Berlin, started its work last year.

"The MPS had certainly been reluctant to dig into the details of its 'dark time'", says Kaufmann. She points out, for example, that only 50 pages in a 1,000-page 'Festschrift' commemorating 75 years of the KWS/MPS relates to the period 1933–45. "It's embarrassing."

Kaufmann insists that "we don't want to



Spot the difference: Werner Köster, handing Max Planck a goblet filled with wine from 1911 — the year that Planck won the Nobel prize. The ceremony took place at the Institute for Metallurgy in Stuttgart in 1934. The Nazi officer was removed in the version printed in 1949 in the *Festschrift* celebrating the institute's twenty-fifth anniversary (left). The original reappeared in a 1989 paper. The discrepancy was spotted by historian Helmut Maier.

uncover one scandal after another, though this will, of course, happen in time". Although keen "to make clear that atrocious experiments were carried out", she adds that "it is important to go further and focus on scientific practice itself and the role of science in shaping policies".

The research programme aims to investigate the role of Kaiser Wilhelm scientists and research institutes in supporting Nazi policy, and the extent to which they took advantage of opportunities that arose as a result of this policy to advance their careers.

The programme also seeks to identify the role of scientists such as the Vordenker der Vernichtung (planners of extermination) in conceiving and developing racial policies, as



Kaufmann: peering into the 'dark time'.

well as in military research and Ostforschung — research in the conquered eastern territories, particularly agricultural and materials research, intended to support Nazi policies of self-sufficiency.

At the specific request of Markl, the historians are also investigating the degree to which scientists of the postwar MPS attempted to call back — or compensate — the 70 or so mostly Jewish KWS scientists who had been thrown out of their institutes during the Nazi period and forced to emigrate.

Although the MPS senate formally recognized such scientists as "External Scientific Members" in 1948, Markl says he would like to know "how intensively and systematically" the society tried to make amends or involve them in the reconstruction of science after the war.

Until the 1980s, says Kaufmann, the dominant view of the role of scientists in the Third Reich — propagated partly by the scientists themselves — was that they had either concentrated on their research and kept out of politics, or been forced to do research sup-

porting Nazi policies to ensure their own survival. Many argued that the latter was not real science, but 'pseudoscience', says Kaufmann.

Recent archival research, and a new willingness to accept that even 'good' science, rigorously conducted on dispassionate and rational criteria, can be morally questionable, have turned this view around, she says. But Kaufmann notes that, even now, many scientists are reluctant to accept that 'good' science was carried out by respected researchers deliberately supporting Nazi policy, and her work sometimes "makes her unpopular" in the scientific community.

From lab to concentration camp

One example of such rigorous, if controversial, research was that carried out by Ernst Rüdin. The director of the Kaiser Wilhelm Institute for Psychiatry in Munich, Rüdin was widely recognized for his theories, published in 1916, on the heritability of schizophrenia. Recent research has shown that he was also instrumental in drawing up the Nazi law on compulsory sterilization of mentally handicapped people, insisting that strict scientific criteria be developed to select those who should be sterilized.

During a six-month study period as guest scientist of Kaufmann's programme, Volker Roelcke, professor of medical history at the University of Lübeck, discovered that Rüdin provided both intellectual and financial support for experiments by Julius Deussen, a young psychiatrist at the University of Heidelberg, designed to develop such criteria.

Deussen used advanced behavioural and biochemical methods to study mentally handicapped children. The children were then killed, their brains examined and the results compared with clinical data. "This was not pseudoscience, but real science, using up-to-the-minute methods," says Roelcke. The aim was to determine correlations between different data sets to aid the selection of children for sterilization. "We never contribute usefully to society."

Such stories are not the whole history of science in the Third Reich. Much of Kaufmann's programme addresses how scientists used the opportunities resulting from Nazi policies and the war. Susanne Heim, for example, a member of Kaufmann's team, is investigating several such cases, including that of Hans Stubbe, who became a top plant geneticist in East Germany after the war.

Stubbe, says Heim, appears to have been an opportunist, ruthlessly following any path that would help his research and career. His strong international connections meant that the Institute for Plant Genetics and Research into Cultivated Plants in Gatersleben near Halle, of which he was director, became well known in the West.

The institute was particularly highly regarded because it housed the second largest plant collection in Europe, after Leningrad. Stubbe made the most of the opportunities created by the Nazi policy of 'autarky', or self-sufficiency in food supplies. During the war, for example, he went to the Balkans immediately after their conquest, accompanying the Italian and German armies, to collect wild plants.

As the first director of the Kaiser Wilhelm Institute for Cultivated Plant Research near Vienna, which opened in 1943, he also collaborated with the SS to plunder valuable Russian collections of wild and cultivated plants after Russia was invaded. He even tried to join the SS himself to make his work easier, but as a known left-winger, his application was turned down.

"Stubbe was not an out-and-out Nazi," says Heim. "It is important to study the relationship that successful KWS scientists had to politics. Scientists like Stubbe were very much aware of the political context in which they worked, even though many scientists like to believe they 'just got on with their basic research.'" **Alison Abbott**

Berlin research said to be threatened by hospital budget cuts

Munich The Max Delbrück Centre for Molecular Medicine (MDC) in Berlin is concerned at plans by the city government to halve the number of beds at two university hospitals at the MDC's campus at Berlin-Buch. The cuts, announced last month, are intended to save the city's public health insurance system DM60 million (US\$30 million). But Detlev Ganten, head of the MDC, warns that such a move would "irreversibly damage" research at the hospitals and the MDC.

Some projects funded by the Deutsche Forschungsgemeinschaft (Germany's research agency), the federal science ministry and the European Union would "collapse", says Ganten. Berlin's medical research community has collected more than 40,000 signatures against the proposed cuts. A decision is expected later this month.

Tennessee gives Oak Ridge neutron source a tax break

Washington The Spallation Neutron Source (SNS), the large neutron facility being built at the Oak Ridge National Laboratory, has been exempted from \$30 million of state taxes,

thanks to special legislation signed by Don Sundquist, the governor of Tennessee.

The levying of Tennessee's 'use tax' on the construction of the SNS had been strongly criticized in Congress, where the exemption is expected to increase the project's chances of receiving adequate construction funds. Jim Sensenbrenner (Republican, Wisconsin), chair of the Science Committee in the House of Representatives and a frequent critic of the project, called the exemption "a big step towards making the SNS project at Oak Ridge a reality".

Genetics remembers its 'cautionary' past

Washington More than 1,200 photographs and other records documenting the American eugenics movement — which promoted the selective breeding of human beings for socially desirable traits — have been placed on the Internet, hosted by the DNA Learning Center at Cold Spring Harbor Laboratory, New York.

Eugenicists in the early twentieth century justified the segregation, limiting of immigration, and sterilization of people considered genetically unfit on what they claimed were sound scientific grounds. The exhibit tells a "cautionary tale" for the coming post-genomic era, says David

Micklos, the project's leader. The exhibit is available at: <http://vector.cshl.org/>

US telescope lends South African project its expertise

Cape Town An agreement was signed last week formalizing the participation of the Hobby-Eberly Telescope (HET), located at McDonald Observatory in Fort Davis, Texas, in the design and construction of the \$16.5 million Southern African Large Telescope (SALT), to be built at the South African Astronomical Observatory in Sutherland, South Africa.

Ben Ngubane, Minister of Arts, Culture, Science and Technology, gave the project the green light to begin construction last November (see *Nature* 402, 452; 1999). The nine-metre-class SALT is based almost entirely on the design of the HET. The five US and German universities making up the HET partnership will provide innovative telescope design, software, commissioning experience, and technical expertise, in exchange for 10 per cent of observing time when SALT begins operation, scheduled for 2003.

Marine biologists axe Ukraine meeting

London This year's annual European Marine Biology Symposium (EMBS) has been

cancelled in protest at the charges brought against marine biologist Sergei Piontkovski. Piontkovski worked at the Institute of Biology of the Southern Seas in Sebastopol, Ukraine, which was to host the symposium in September (see *Nature* **402**, 6; 1999). He was arrested last November on charges of sending secret information abroad and illegally handling foreign currency.

After a poll of the EMBS advisory committee, EMBS president Giulio Relini, from the University of Genoa, said that "European researchers wish to give a strong and clear message of their disappointment and unhappiness about the incredible story of Professor Piontkovski". No replacement symposium will be organized.

Judge ponders next move in *Taq* patent case

San Diego US District Judge Vaughn Walker last week set 24 February as the date for hearing arguments relating to his December ruling that the Swiss company Hoffman-LaRoche secured their patent for natural *Taq* polymerase, an enzyme used in the polymerase chain reaction (PCR), fraudulently (see *Nature* **402**, 709; 1999).

Promega Corp. of Wisconsin will argue before the judge in San Francisco that two of Roche's patents for the PCR process are also

fraudulent, and contain some of the same misrepresentations used in the patent for natural *Taq*.

Promega, which sells natural *Taq*, will seek to have the patents invalidated. Hoffman-LaRoche denies any fraud, and will seek to appeal against Walker's ruling, accusing him of "a series of scientific errors".

Robot goes it alone on meteorite hunt

London This four-wheeled robot has been deployed in a remote part of the Antarctic to search autonomously for meteorites and classify them in the field. The robot, called Nomad, is a joint venture between researchers from Carnegie Mellon University's Robotics Institute and the US National Science Foundation's Antarctic search for meteorites programme.

The robot uses high-resolution imagery



and spectroscopy to gather data about the rocks it finds. After only three days of searching in an area 160 miles northwest of the US base at McMurdo Station, Nomad has already found one rock that it has classified as a meteorite, and another seven classified as terrestrial.

Topologists and neutrino hunters win Wolf prizes

Jerusalem Raoul Bott, of Harvard University, and Jean-Pierre Serre, of the Collège de France, Paris, share the annual mathematics prize awarded by the Wolf Foundation. Bott was cited for his deep discoveries in topology and differential geometry and their applications to Lie groups, differential operators and mathematical physics. Serre was cited for his fundamental contributions to topology, algebraic geometry, algebra and number theory, as well as for his inspirational lectures and writing.

Raymond Davis Jr, of the University of Pennsylvania and the Brookhaven National Laboratory, and Masatoshi Koshiba, of the University of Tokyo, share the physics prize for their pioneering observations of astronomical phenomena by detecting neutrinos, which created the emerging field of neutrino astronomy. Each prize is worth \$100,000.

Apathy rewards misconduct — and everybody suffers

Sir — Louis Guenin's Commentary (*Nature* **402**, 577; 1999) is a welcome contribution to defining misconduct in scientific research. A related problem is that research misconduct is all too frequently seen as a victimless crime to which indifference is an adequate response. As recently as October 1997 the heads of the UK research councils were reported in *Research Fortnight* to have decided that misconduct is a lesser evil than the encumbrance of any mechanism to prevent it.

More recently, the research councils have had a change of heart and published policies on misconduct. But there is no mechanism for ensuring compliance, so institutions can whitewash misconduct or sweep it under the carpet. Since investigations bring adverse publicity, these are tempting options. Nor is there any adequate means of protecting honest whistle-blowers — a gaping hole, given the fraudster-friendly nature of UK libel law.

All honest scientists are victims of scientists who commit misconduct. Jobs in science, research funds and journal space are all scarce. Every job occupied, every grant received and every paper published by someone who engages in misconduct deprives at least one honest scientist of an opportunity to which he or she was entitled. To that can be added the waste of time and resources when other scientists attempt to use fraudulent or misrepresented results.

Scientific fraud resembles financial fraud in that it can bring undeserved remuneration and power, a salient difference being that in scientific fraud the ill-gotten gains are automatically institutionally laundered.

Herbert N. Arst Jr

Department of Infectious Diseases, Imperial College School of Medicine, London W12 0NN, UK

End of impact factors?

Sir — Does the expansion of the Internet and journals' online publishing strategies herald the end of 'impact factors'? Many scientific publications now have online versions, often freely available. As Internet access is pervasive (and increasing), busy scientists can now sit with their morning coffee and use a computer search engine to look for articles in their narrow area of interest. The computer database can be much more exhaustive, user-friendly and up-to-date than its old-fashioned paper counterpart. As a result, the researcher can now in effect have a customized

'table of contents' generated on demand.

The overall effect of this practice seems likely to be a shift in reader emphasis away from particular ('high-impact') journals as reference sources, and an increasing importance of specific articles, rather than the journal in which they are published. If this is indeed the case, continued emphasis on 'impact factors' as currently calculated would seem to be misguided, and the concept will need to be redefined.

John Brunstein

Department of Virology, Haartman Institute, University of Helsinki, Haartmaninkatu 3, 00290 Helsinki, Finland

Nucleic acids revelation delayed by a sceptic

Sir — Although the discovery of nucleic acids, mentioned in Peter Little's News and Views article "The book of genes" (*Nature* **402**, 467; 1999), was published in 1871, the Swiss physiologist Johann Friedrich Miescher (1844–95) actually made the discovery two years earlier.

Working in Tübingen between autumn 1868 and 1869, Miescher isolated the new substance from cells of pus (leukocytes) obtained from discarded bandages from the local surgical clinic. In order to find the chemical constitution of nuclei, he removed the proteinaceous cytoplasmic substances of the cells by digestion with gastric juice (containing the protease pepsin) from pigs' stomachs and with hydrochloric acid. The material of the naked nuclei that he obtained and called "nuclein" contained 14 per cent nitrogen, 2 per cent sulphur and 6 per cent phosphorus pentoxide, making it very rich in phosphorus.

The final studies were made in the autumn of 1869, because on 21 August that year Miescher wrote to his parents: "I still have to complete the definitive analysis of the nuclear substances." And the paper is dated "Basel, October 1869". Miescher's teacher and laboratory chief in Tübingen, the German biochemist Felix Hoppe-Seyler (1825–95), received the manuscript, but was sceptical about the rather revolutionary findings of a beginner. Therefore, he decided to repeat the experiments, and he printed Miescher's paper "Über die chemische Zusammensetzung der Eiterzellen" only after he had verified them.

Hoppe-Seyler wrote in a footnote that the publication was delayed very much "through several unforeseen circumstances". In a paper of his own he writes: "I have to emphasize that in all points as far as I have examined Miescher's statements I have to confirm the latter fully."

Friedrich Katscher

Mariahilfer Str. 133, A-1150 Vienna, Austria

Proteomics is getting easier in some ways...

Sir — The report on proteomics (*Nature* **402**, 715; 1999) was an excellent summary of current attempts to build on knowledge of the genomes. I would point out, however, that 2D gel electrophoresis is no longer "notoriously difficult to carry out". My undergraduates routinely perform 2D gels, often getting ten good gels on the first try. Also, while hydrophobic proteins present a challenge, some of them can indeed be obtained on gels using SDS/urea treatment of the sample; this process is compatible with current gel technology.

Joan L. Slonczewski

Department of Biology, Kenyon College, Gambier, Ohio 43022, USA

... and should be treated the same as genomics

Sir — Your Briefing on proteomics was valuable and informative. Unfortunately, the Opinion (*Nature* **402**, 703; 1999) was less so. The writer notes that "*Nature* intends to play its part by insisting on conceptual insights from among the great quantities of information that [proteomics studies] will certainly deliver".

I disagree with requiring such standards for studies of proteins when very different standards appear to be applied to sequencing studies. For example, the same issue contained two articles describing the sequencing of chromosomes 2 and 4 from *Arabidopsis thaliana*. As a researcher from outside the plant field, I did not find conceptual insights in this information. The speculations regarding mitochondrial gene exchange and relative proportions of receptor-signalling proteins were interesting, but were similar to speculations that would arise out of most proteomics studies where specific subsets of proteins would be identified for particular cell types.

This is not to say that I think publishing genomic sequencing milestones in *Nature* is inappropriate. Rather, I think you should publish both sequencing and proteomics studies, and apply similar standards for evaluating them. I find it likely that proteomics studies will contain inherently more conceptual insights, since the proteomics will be able to use the genomic sequencing information to make correlations between expression patterns and promoter and other sequence information.

Jeffrey E. Segall

Department of Anatomy and Structural Biology, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, New York 10461, USA

The decade of the sheep

How a discredited technique led to the potential for creating new species.

The Second Creation: The Age of Biological Control by the Scientists that Cloned Dolly

Ian Wilmut, Keith Campbell
and Colin Tudge

Headline: 2000. 362 pp. £18.99

Anne McLaren

Not many books describe how science is actually done. James Watson's *The Double Helix* was one, and this book is another. Although both books are fascinating, they are interestingly different. James Watson and Francis Crick are very different people from Ian Wilmut and Keith Campbell. When writing his book, Watson had literary guidance, including that of the novelist Naomi Mitchison; in writing their book, Wilmut and Campbell have the accomplished science writer Colin Tudge as co-author. *The Double Helix* is set against the background of an American graduate student's life in Cambridge in the early 1950s, and focuses on months, weeks and days of intense intellectual activity. The background of *The Second Creation* is UK government-funded research over the past 30 years. Decades of goal-oriented experimentation are described, with all the inevitable frustrations, errors and blind alleys, but also 'eureka' moments of great excitement and achievement. Finally, if Crick and Watson had not cracked the structure of DNA when they did, someone else would have done so within a year or so, if not sooner; but without Wilmut and Campbell, mammalian cloning from differentiated cells might have been delayed for a decade or two, because nobody else was working on it seriously.

The casual reader might assume that 'the second creation' refers to the creation of Dolly, the sheep cloned in 1996 from an adult cell by Wilmut and Campbell at the Roslin Institute in Scotland. Not so. The book focuses on the coming together of three technologies: genomics, genetic engineering and cloning by nuclear transfer from cultured, differentiated cells. This century, genomics will make it possible to map all the genes of any animal and find out their DNA structure; genetic engineering will allow any DNA sequence to be precisely added, removed or replaced, using homologous recombination in cultured cells; and cloning by transfer of nuclei from cultured cells into enucleated, unfertilized eggs means that the immense power of genomics and genetic engineering can in principle be applied to animals. New varieties and even new species could be created at will — hence the second creation.

At the Roslin Institute, the story really



Ewe too: Dolly was just one of a whole series of genetically transformed sheep produced at Roslin.

began in 1990 with the birth of Tracy. Tracy was the culmination of a research project aimed at producing genetically transformed sheep that would express valuable pharmaceutical products in their milk, using the well-established but relatively inefficient technology of injecting DNA constructs into the pronuclei of fertilized eggs. By this means, genes can be added but not removed or replaced; they are inserted at random, may be expressed only at low levels or not at all, and are not always transmitted to the next generation. But Tracy and her offspring produced high levels of α 1-antitrypsin (a drug used to treat cystic fibrosis and other human diseases) in their milk.

Tracy was a triumph, although the principle had already been established in mice. Then, in 1995, Megan and Morag were born, arguably the most important elements in this whole series of sheep. Megan and Morag were cloned by transfer into enucleated unfertilized eggs of nuclei, not just from early embryonic cells (that had been done before), but from early embryonic cells that had been put into short-term culture and had differentiated. So a nucleus that had started down one pathway of differentiation could be reprogrammed by egg cytoplasm to give rise to a whole adult animal. Although this had been achieved in frogs, it was widely thought to be impossible in mammals. Media attention was less than might have been expected, perhaps because the massacre of 16 schoolchildren in Dunblane, Scotland, occurred immediately after.

The following year saw the birth of Taffy

and Tweed, cloned from nuclei of differentiated fetal fibroblasts put into long-term culture (not possible with early embryonic sheep cells), and also Dolly, the first adult animal cloned from the nucleus of an adult cell (never achieved in frogs). Dolly caught the world's attention because of the possible human implications, but from the Roslin point of view she was just the gilt on the gingerbread. In 1997 Polly was born, cloned from a long-term cultured fetal fibroblast nucleus into which had been randomly inserted the human gene for blood factor IX (used to treat haemophilia). Rumour has it that PPL, the commercial offshoot of Roslin, has now produced Diana and Cupid, cloned from genetically targeted fetal fibroblast nuclei, and producing human α 1-antitrypsin in their milk — like Tracy, but stemming from an immensely more powerful technology, all developed within a decade.

Wilmut and Campbell are complementary people and played complementary roles. Wilmut is an agricultural scientist who wanted to use genetic engineering to improve farm animals. In 1987 he had a sudden flash of insight. Cloned lambs had recently been produced by nuclear transfer from the inner cells of early sheep embryos; in mice, these same cells could give rise to embryonic stem cells that could be genetically manipulated; so if sheep embryonic stem cells could be made and their nuclei used for cloning, his ambition could be realized. But making sheep embryonic stem cells proved a problem. Campbell is a cell scientist, who had worked

on cancer, cell differentiation and the cell cycle. He believed that differentiation was not irreversible, and his insight that nuclei from differentiated cells could be reprogrammed by egg cytoplasm, if only the cell cycle were better understood, provided the alternative to embryonic stem cells and made possible Megan and Morag, Dolly and Polly.

A fascinating and comprehensive account of the history of cell biology and embryology over the past 150 years puts this story in context. Rudolf Virchow proposed in 1855 that "all cells come from cells"; August Weismann in the 1890s believed that all development and differentiation consisted of loss of "hereditary material"; Wilhelm Roux claimed to have experimental support for Weismann's theory, but Hans Driesch disproved it by splitting sea-urchin embryos into four separate cells at the four-cell stage, each of which gave rise to a complete embryo — the first animal cloning experiment.

In the early years of the twentieth century, first Jacques Loeb in sea urchin and then Hans Spemann in salamanders, used a primitive form of nuclear transfer to show that nuclei retained totipotency during cleavage and could support development of whole embryos if returned to egg cytoplasm. Inspired by these results, Spemann in 1938 proposed a "fantastical experiment", namely to take a nucleus even from an adult cell and place it in the cytoplasm of an egg whose own nucleus had been removed. He saw no way of doing the experiment, but appropriate techniques were later developed in amphibia by Robert Briggs and Thomas King in the 1950s and by John Gurdon in the 1960s. Gurdon's cloning experiments showed that even fully differentiated cells could regain totipotency if exposed to egg cytoplasm, so no loss of "hereditary material" had occurred.

The success of Gurdon's amphibian experiments inspired several cloning-based novels in the 1970s, as well as an interest in mammalian cloning. *The Second Creation* gives a detailed and sympathetic account of the 1981 claim by Karl Illmensee and Peter Hoppe to have cloned mice by transfer of nuclei from the inner cells (ICM) of early embryos. Systematic attempts to repeat their work, notably by James McGrath and Davor Solter, failed totally. In 1984 McGrath and Solter (in a *Science* paper whose title is, unfortunately, misquoted) concluded not only that "ICM cell nuclei are unable to support development with enucleated zygotes", but also that "the cloning of mammals, by simple nuclear transfer, is biologically impossible". Solter is a highly respected developmental biologist, and so, on the 'authority principle', many people were deterred by this statement from pursuing mammalian cloning further.

The section on the history of Roslin is a microcosm of the chequered history of scientific research in Britain. In 1982, the

Agricultural Research Council (as it then was) was in financial crisis, and planned to close the Animal Breeding Research Organisation (as Roslin then was), but decided instead to cut its funding drastically — I can vouch for this; I was on the council at the time. Wilmut nearly lost his job; in the re-organization he was directed to change his line of work — but this turned out rather well.

In spite of its length, and a certain amount of repetition, this book is very readable. Non-scientists will not only learn a lot of biology, but they will also get a good idea of what makes scientists tick. There is a useful glossary and an adequate index, but almost no references are given. This is a major omission: many readers, including myself, will want to follow up some of the work described.

One of Wilmut's motivations when he started to work on embryos was that he "found them very beautiful". He says he "would like to be appreciated". Campbell says, "I don't want fame; but... some recognition would be nice". Scientists will warm to the constant emphasis on the importance of serendipity, the importance of meeting people, of travel, conferences, of just talk. The book is excellent value for money: if any biologically minded teenager decides to aim for a career in research as a result of reading it, I think the authors will be well-pleased. ■

Anne McLaren is at the Wellcome/CRC Institute, Tennis Court Road, Cambridge CB2 1QR, UK.

Dead branches on the tree of life

In Search of Deep Time: Beyond the Fossil Record to a New History of Life

by Henry Gee
Free Press: 1999. 304 pp. \$26

David M. Irwin

What is deep time, and what does it have to do with the fossil record? Henry Gee describes it as the impenetrable depth of time that prevents us from ever knowing in true detail the evolutionary history of our own or any other species. The written history of the past few thousand, or even hundred, years is not absolute, so how can we know with certainty events that occurred millions of years ago? But all is not lost, Gee explains, because an objective scientific method exists which has illuminated many events in the history of life, and has shown that they are often stranger than we commonly believed.

The goal of *In Search of Deep Time* is to destroy several assumptions that are held by most people — including biologists — so that we can rebuild our view of the

history of life. Remember that although a living individual must have had ancestors, fossils are unlikely to represent any of them. Even if a fossil was an ancestor, we will never know this — we can never know with certainty what happened in the past. Accepting that fossils are not ancestors also means that there are no 'missing links' in the fossil record because fossils cannot be ordered, as traditionally depicted, into an evolutionary lineage. There is no ladder of life. Most, if not all, fossils lie on the dead branches of the tree of life, and we must remember that most of our tree of life is dead, with only a few green living leaves at the tips of the branches.

What we do know, though, is that fossils are from species that are related to us. We need to determine how they are related to us, and with this phylogeny we can begin to imagine what our ancestors were like. How do we determine these relationships? Which of the many extinct species found in the fossil record are we more closely related to, and how are all these extinct species related? Gee explains how cladistic analysis can be used to answer these questions, and illustrates this with several examples that have unearthed relationships never before imagined.

Gee presents a clear explanation of the principles of cladistics, starting by trying to determine the relationship between himself and his two cats, Fred and Marmite. Technical details are ignored, but the power of cladistics is well demonstrated. From this basis the book illustrates how cladistics allowed hypotheses to be developed which could be tested with data — something that was impossible for previous evolutionary explanations proposed by 'authorities'. Gee explains, from personal experience, how cladistics entered the field of palaeontology — of deep time — the resistance to its use, and how it has revolutionized our interpretations of fossils.

With the realization that most fossils are not the ancestors of any living species, indeed, that the tree of life is mostly made up of dead branches and that fossils come from these branches, it is time to destroy our second assumption — that all life must resemble what we see today.

Our view of life and the Universe is constrained by our own experiences — our imagination is limited. How many digits did the limbs of the first tetrapod have? Most living descendants have five — so was it five? Wings are used for flying, so did wings evolve for flying? For both questions, the examination of fossils and cladistic analysis tell us that the answer is almost certainly no; as Gee explains, we probably will never know the true answers. This also leads Gee to ask whether we really know what to look for when searching for life outside this planet.

If there are some faults with the book they don't detract from the main message — that fossils are not ancestors and we know very little about our evolutionary history. The author implies that cladistics is the best method for solving many of the problems of phylogeny. While I agree that cladistics is powerful, other approaches are also useful. It was not just through cladistics that scientists came to believe that hippos are more closely related to whales; other approaches, such as distance methods, yielded the same conclusions. Cladistics should not become the new 'authority'.

As *In Search of Deep Time* is written by a senior editor at *Nature*, the author has been in a position to have seen many of the original manuscripts — and reviews of them — that describe recent advances in our understanding of the evolution of life on Earth. Gee's comprehensive knowledge is clearly evident in his book. He has also added his personal experiences, and imaginings, in East Africa and in the galleries and back rooms of the Natural History Museum.

This book is intended for the informed reader, and as such is to be recommended. As an evolutionary biologist I did not need to be told that many assumptions held by most people are false, yet I think I needed to be reminded that my imagination is limited.

We don't know what our ancestors really looked like.

David M. Irwin is in the Department of Laboratory Medicine and Pathobiology, Faculty of Medicine, 100 College Street, University of Toronto, Toronto, Ontario M5G 1L5, Canada.

Textiles, bridges, plastics and chips

Technology in America: A Brief History

Alan I. Marcus and Howard P. Segal
Harcourt College: 1999. 400 pp.
\$43, £14.95 (pbk)

Gunhard Æ. Oravas

It is a formidable undertaking to write a history of technology that takes full account of cultural, economic, social, political and scientific influences, especially if the history is to be concise as well as precise. This superbly composed history of technology is a result of many decades of scholarly effort. Although the book is loaded with information and explanation, the authors make the technological topics under discussion interesting to the general reader by frequently referring to well-known historical figures.

Their history begins with the colonial period in 1607, when most of the industry consisted of textile manufacture, leather-making, iron production and shipbuilding. In the early nineteenth century, the building of bridges, canals, steamboats and railways and the use of machinery in the textile industry became prominent. The latter part of the nineteenth century saw a mass-production steel industry, the telegraph, telephone, electricity, the organization of factories and engineering and the systematization of technical education.

In the twentieth century, the founding of industrial research laboratories, the development of plastics such as Bakelite and Nylon, the establishment of new delivery technologies such as lorries and public transport after Goodyear tyres were marketed in 1916, and the emergence of commercial aviation as a serious competitor in delivery and travel are delineated with precision.

Of the many entries of interest to computer enthusiasts and general readers, Alan Marcus and Howard Segal's discussion of the development of the programmable microchip, and of integrated-circuit technology and software will be of special note.

The authors have also not overlooked such very recent developments as biotechnology, genetic engineering, cloning, ultrasound, magnetic resonance imaging and computer axial tomography. This history of American technology is a work of great merit which is written very skilfully in an intelligible form that is also easily accessible to the educated general reader.

Gunhard Æ. Oravas is in the Department of Civil Engineering & Engineering Mechanics, McMaster University, Hamilton, Ontario L8S 4L8, Canada.

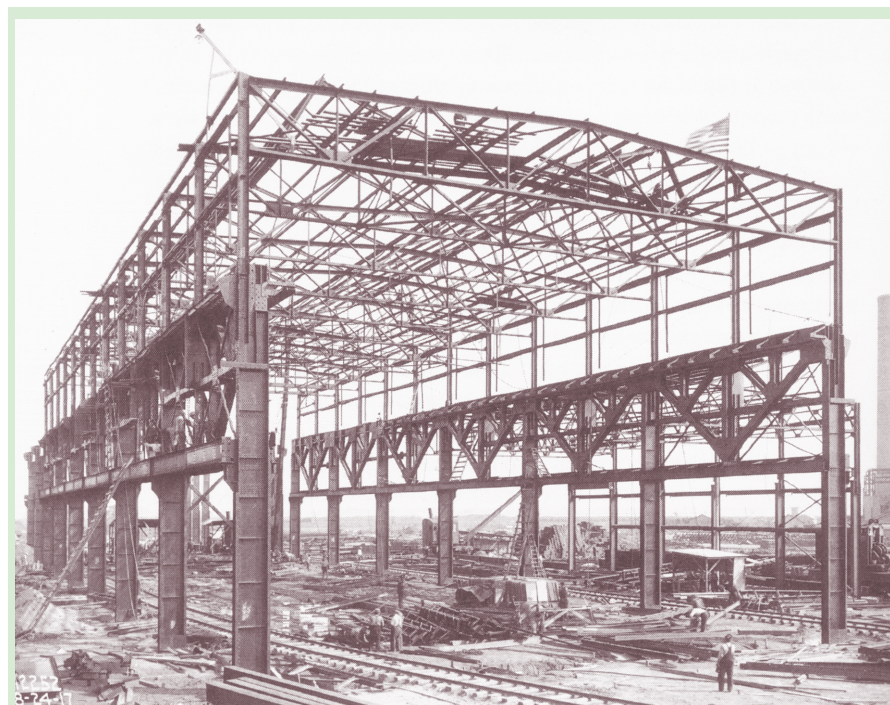
Cocktail-party gossip aplenty

Primate Sexuality: Comparative Studies of the Prosimians, Monkeys, Apes, and Human Beings

by Alan Dixson
Oxford University Press: 1998. 656 pp.
£32.50, \$60 (pbk)

Jeanne Altmann

During several decades of fieldwork on natural primate populations, I, like many others, have learned that children in tow are a valuable entrée in most parts of the world. Almost as good as an entrée, however, and of considerably more interest at a cocktail party, is the opportunity for people to satisfy their considerable curiosity about the sex lives of our closest relatives. In *Primate Sexuality*, much about primate sex is revealed. Perhaps more is revealed than all but the most devoted even realized they might want to know



Building industrial art

Developments in factory buildings and works during the nineteenth and early-twentieth centuries reflect a subtle relationship between function and aesthetics. This aspect of the industrialization of the United States is traced in

Betsy Hunter Bradley's *The Works: The Industrial Architecture of the United States* (Oxford University Press, £37.50, \$45). Above is a 1919 construction photograph of the foundry building of the Westinghouse Electric Company.

about primate sexuality, but not more than many will find of interest. That the book has a foreword by the director of the Kinsey Institute is hardly surprising.

Alan Dixon has drawn together an impressive body of literature in this landmark volume. The book takes a comparative perspective, seeking a seamless review from prosimians through humans, and viewing this taxonomic diversity from a Darwinian and phylogenetic perspective. Moreover, although Dixon indicates that the book's focus was intended to be behavioural, it also devotes much attention to the morphological and hormonal aspects of the behaviour considered. As a result, Dixon's treatment is unusually inclusive across levels of analysis and activity. His reach is impressive, and he has sought to distil and review more than 2,000 sources from a diverse literature.

The book succeeds admirably, although its success is more modest and uneven than its reach. This is perhaps inevitable, given the fact that it is a first edition and has such ambitious goals. Coverage and quality of analysis too often reflect not just the state of the field but also the author's direct or close exposure — geographically, taxonomically and topically. For many topics, Dixon points out disagreements and competing theories and conclusions in the literature, invaluable for those who otherwise would not even know a debate exists. When he takes a stand on a particular point, however, the basis for that stand is too often not obvious or compelling. This is especially unfortunate in situations where many in the field would disagree with Dixon's reading of the weight of the theory or data.

The book is enriched with many comparative tables, most taken directly or almost directly from other publications. There are also a large number of illustrations and data figures from the literature, often redrawn by the author, which tellingly bring to life particular points and examples in the text. Some tables and figures are less successful and valuable than others, sometimes frustrating the author's comparative goals. Future editions, and surely the book will warrant these, will benefit greatly if comparative tables taken from the literature are edited to remove redundancies and resolve contradictions, and if new tables are produced to show the current state of knowledge.

Primate Sexuality is an essential starting point in this field, and a must for every primatologist's library. Its shortcomings should only provide an impetus for the next edition, which will be necessary if for no other reason than the current rapid developments in the field. ■

Jeanne Altmann is in the Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544, and the Department of Conservation Biology, Chicago Zoological Society/Brookfield Zoo, Brookfield, Illinois 60513, USA.

Science in culture



Constable's paradoxical rainbow

Why the artist John Constable allowed himself poetic licence when painting rainbows.

John E. Thornes

John Constable's powers of observation and thirst for meteorological knowledge enabled him to paint more realistic skies than any other English artist (see *John Constable's Skies*, University of Birmingham Press, 1999). Constable believed that "painting is a science, and should be pursued as an enquiry into the laws of nature. Why then may not landscape painting be considered as a branch of natural philosophy, of which pictures are but the experiments?" This Baconian view was certainly applied to the clouds and weather in his paintings, but, as we shall see, not to all of his rainbows.

Constable's paintings of rainbows are almost as well known as his paintings of clouds, but whereas clouds are present in some form almost every day in England, rainbows are seen much less frequently and are therefore more mysterious. Constable painted rainbows that were meteorologically accurate, but also used them for artistic effect in situations in which they could never occur in nature, such as in *Salisbury Cathedral from the Meadows* (1831) (see above). Although Constable knew that a rainbow cannot be seen once the Sun is higher in the sky than 42 degrees and that the Sun must be directly behind the observer of a rainbow, it plainly is not so in this view of the cathedral. This enigma is difficult to explain. Why should Constable be such a perfectionist about the weather in his paintings, his dutiful wind and clouds and harmonic daylight, and yet be content to introduce a meteorologically impossible rainbow?

Constable wrote to his friend John Fisher in 1821: "The sky is the source of light in nature — and governs everything ... Their difficulty [clouds] in painting both as to composition and Execution is very great, because with all their

Constable's Salisbury Cathedral from the Meadows (1831).

brilliance and consequence they ought not to come forward or be hardly thought about in a picture — any more than extreme distances are — But these remarks do not apply to phenomenon — or what the painters call accidental Effect of Sky because they always attract particularly."

Thus the rainbow, and other optical effects such as crepuscular rays, were given a special exemption by Constable, and although a rainbow cannot 'come forward' in a picture, Constable makes it clear that if he introduces such an effect he wants it to be attractive and 'thought about'.

If one looks at the shadow of the fence-post in the bottom left-hand corner of the picture, the geography and the illumination of the cathedral suggest that the Sun has reached an angle just north of west, which, in mid-August, would be between 6 and 7 p.m. At that time, the height of the Sun in the sky is about 15 degrees, which means the rainbow should have a height of about 27 degrees, significantly smaller than observed. The rainbow is not the bow the artist would see but one that would be seen from the meadow on the right. It is not in the plane of the picture but comes out of the picture at a shallow angle. The lighting suggests that the Sun is to the right of the picture and not, as it should be, behind the observer. It would appear that Constable was quite happy to use an amount of poetic licence for his "accidental Effect of Sky".

Contrast Constable's *London from Hampstead Heath with a Double Rainbow* — painted a few months after *Salisbury Cathedral from the Meadows* was exhibited at the Royal Academy — which showed that he was fully aware of the colour reversal and the size of the secondary bow. ■

John E. Thornes is in the School of Geography, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK.

Decline of the generalist

The vigour of every discipline depends on people of broad vision.

Frederick Seitz

During the 70-odd years of my professional life, I have witnessed a continuously increasing degree of specialization in major areas of cultural activity. While the growing chasm between science and the humanities may once have provided the most obvious example — a matter that concerned C. P. Snow a generation ago — the fragmentation process has extended relentlessly into what were once well-integrated fields. Few scientists under 50 are familiar with, or express much interest in, areas of research outside their immediate professional concern. Specialization within the humanities has become so pronounced that aberrant concepts, such as Black Athena, which is based on the claim that the achievements of classical Greece originated in Africa, can flourish as essentially independent, free-floating intellectual structures.

This trend is partly a result of the growing complexity of most fields of research. Conditions of intense competition leave relatively little time for scholars to cultivate new, diverse interests. But a more important source lies in the changing policies of our educational institutions as they deal, necessarily, with larger numbers of students with narrow ranges of interest.

Most students are not being prepared to become broad, experienced leaders in a highly professional area. Instead they aim to find a useful place in a socio-economic structure that has less and less demand for relatively unskilled manual workers. Alongside this are the distractions of a highly intrusive popular culture, devouring a student's time that could, under other circumstances, be devoted to more intellectual pursuits. In the United States, education in mathematics has been degraded, as have other 'requirements' such as languages or 'core courses' that were standard when I was a student and important in my own development.

Is it desirable that we have a significant group of generalists in all cultural fields? I believe the answer is an emphatic 'yes'. One reason is to nurture progress in every area, because the major steps in the evolution of a field have generally been governed by an oligarchic assembly of experts with much broader than average vision. Even granting that most major advances are initially the work of a gifted individual, the collective opinion of the peer group, whose members have almost invariably made major contributions to the field as a whole, give the endeavour a sense of unity, as well as

guidance to the majority working in it. Individuals such as Albert Einstein, Niels Bohr, Arnold Sommerfeld, Erwin Schrödinger, Werner Heisenberg, Wolfgang Pauli, Paul Dirac, Eugene Wigner and Enrico Fermi, acting loosely as a group, steered the physics community into the age of quantum mechanics. Similar groups of leaders have played comparable roles in quite different areas outside science, such as the arts, history, archaeology, economics and social science.

Barring the guidance of such highly motivated and broadly based leadership, most fields of cultural development will either drift towards dull mediocrity or degenerate into uncoordinated islands governed by individuals of minor stature possessing narrow, idiosyncratic viewpoints. In this respect, the natural sciences have the great advantage that a firm intellectual foundation emerges from the interpretation of accurately reproducible experimental observations of the natural world.

I recognize that excessive conservatism

Few scientists under 50 express much interest in areas outside their immediate professional concern.

may slow or forestall inevitable evolutionary advances. But when this occurs, the stresses within the professional community lead to a reshuffling of recognized authority. For example, the French impressionists struggled to gain deserved recognition.

What is the remedy for the ever-increasing degree of specialization? At the higher levels of education, the forces currently at work, both internal and external, favour specialization. Moreover, many social science and humanities groups at universities, particularly in the United States, are in the grip of movements gathered under the banner of political correctness, which rejects all or most of the patterns of authority that guided professions in the past. So reform and renewal must begin at the elementary and secondary levels of education.

Schools must recognize that 'élite' students can absorb a far more diverse programme of material than the average student. Here one deals with flexible, unconstrained minds capable of crossing disciplinary boundaries at will and possessing the originality and freedom to do so. Faculty and parental guidance at an early stage may help the process. Many generalists of the future would emerge naturally from this. The teachers involved must be professionally prepared in the domains they cover and dedicated to their mission. In general, they will merit special monetary compensation. ■

Frederick Seitz is at the Rockefeller University, 1230 York Avenue, New York, New York 10021-6399, USA.



Tunnel vision: changes in society have driven higher education to ever greater specialization.

From Caribbean to Clementine

The road to the stars begins in the ocean depths.

Stephen Baxter

It is the year 2020. A near-Earth asteroid, a flying mountain of water ice and organics, is approached by the Clementine, a mining craft from Earth. The asteroid's riches are expected to transform the planet's economy.

As the craft gently nuzzles against coal-dark asteroid dust, the pilot signals success to her mission controllers on Earth. This is an astronaut who can navigate through space, manipulate her environment and control complex machinery. But she is no human. She is a member of the species *Sepioteuthis sepioidea*: a Caribbean reef squid — or rather, a genetically enhanced descendant of natural squid.

Her streamlined, torpedo-shaped body is a rich burnt-orange, mottled black. Wing-like fins ripple elegantly alongside the body. Her head is crowned by a beak and surrounded by flipper-like arms, and there are two forward-looking eyes, blue-green rimmed with orange.

Alien eyes. Intelligent.

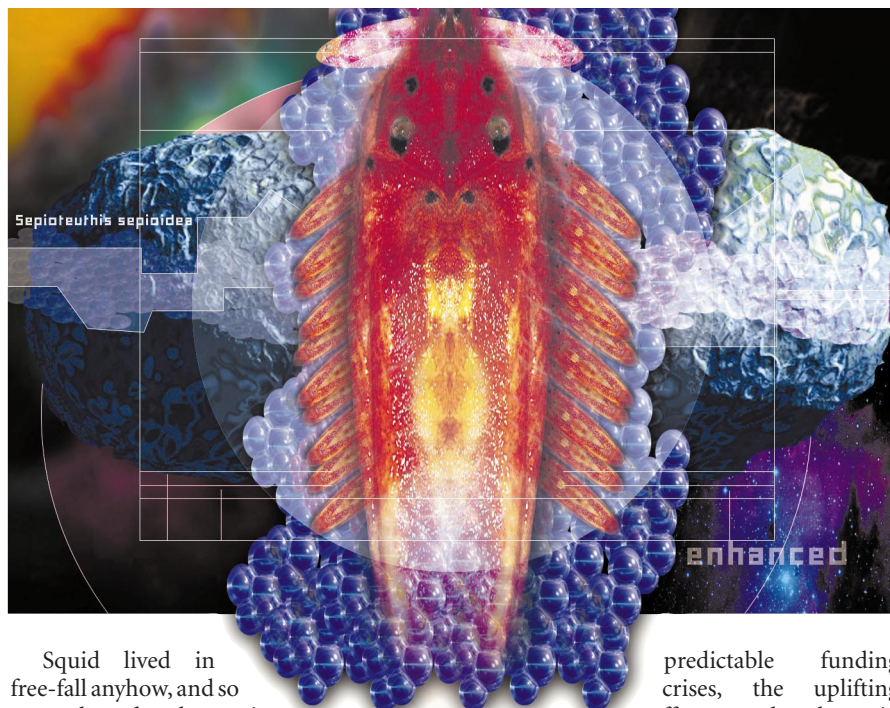
The notion of squid in space would have seemed fantastic to followers of the first human astronauts. But the logic that had led this first cephalopod pilot from the Caribbean to the Clementine had proven inescapable.

It had been understood for some decades that squid were extremely smart molluscs, functionally equivalent to fish. They were highly efficient predators, able to navigate and hunt in three dimensions. They had arms capable of manipulation, used, for example, during mating. They had senses based on light, scent, taste, touch, sound, gravity, acceleration, perhaps even an electric sense.

Squid spoke to each other. Their hides sparkled with patterns made by sacs of pigment granules surrounded by muscles. But were the signals a true language? Human scientists had been able to isolate a number of cephalopod linguistic components that combined in a simple grammar.

But nobody knew what the squid might talk about. They shoaled for mutual protection, but didn't play or groom, and hunted individually. It wasn't clear why such short-lived, only partially social animals needed such complex communication systems.

Nevertheless, in the early twenty-first century, laterally thinking space scientists realized that cephalopods — in particular the reef squid — could have precisely the toolbox needed to equip them for uses in space.



Squid lived in free-fall anyhow, and so were adapted to the gravity-free conditions of space — for them there would be none of the physiological problems that plagued human astronauts. Their life-support requirements were simple: a bubble of water with a basic oceanic ecosystem would suffice. They could navigate with their powerful, predators' eyes. They could communicate. And they could manipulate their environment.

A major international project — funded by NASA, ESA and various oceanographic institutes — was initiated. Its goal was nothing less than to enhance the intelligence of reef squid.

It was a major technical challenge. A squid's neural layout is not like that of a human. A squid has two nerve cords running like rail tracks the length of the body, studded with pairs of ganglia; the most anterior pair of ganglia are expanded into a mass of lobes. The scientists quickly identified the areas of the brain responsible for language learning (unpleasantly, by cutting away parts of squid brains to see what happened.) It was discovered that some comparatively simple genetic engineering could be deployed to make squid smart — or rather, smarter.

There were protests, of course, from animal-rights activists: but also anguish of a more philosophical kind, from groups that questioned the ethics of our ability to inflict an awareness of mortality on another creature not previously sentient.

Despite such opposition, and also the

predictable funding crises, the uplifting effort produced rapid results. Soon, enhanced squid were performing useful work in the ocean — working on sea farms, for example. And the first experiments were performed to see if they could operate outside the Earth altogether.

To no one's great surprise, they made fine astronauts.

Eerily, as some pointed out, it was almost as if the squid had evolved for the conditions of space travel — as humans self-evidently hadn't. Some conspiracy theorists began to wonder if the primary purpose of humanity had been, all along, to deliver enhanced cephalopods to their natural environment in space. And others, wary of the squid's huge potential, began anxiously to watch the skies.

She can feel the feather-touch of new gravity. Beneath the translucent skin of her habitat she can see a grainy, grey-black ground, a jagged horizon, barely tens of metres away. Earth is very remote, and her ties of loyalty to humanity are stretched. To the squid pilot, the warmth in her mantle cavity — impregnated eggs — is much more important.

And to her, this asteroid is no mine in the sky. It is a breeding ground. She flashes her triumph, her mantle skin tingling. At last, with a sense of excitement, she slips her arms into the waldoes and prepares to begin her work.

Stephen Baxter writes science fiction that is extrapolated from current understanding. His latest novel is *Time* (HarperCollins).

Fluff balls of fire

Graham K. Hubler

The most mysterious sort of lightning is ball lightning — glowing spheres of light that float in air. A new theory claims to explain nearly all the properties of these unusual balls of fire.

Ball lightning has been well documented since the Middle Ages as a natural phenomenon associated with thunderstorms. It is relatively rare — only about 1% of the population ever reports seeing it (Fig. 1). It remains an enigma to modern science. Experiments to reproduce ball lightning in the laboratory have not been successful, and a theoretical explanation has eluded scientists since the first attempts 150 years ago¹. On page 519 of this issue, Abrahamson and Dinniss² offer a straightforward explanation of ball lightning, in which they suggest that fluffy balls of silicon burn and emit light. The balls of silicon are created when ordinary fork lightning strikes the earth, by the same chemical reaction that the semiconductor industry uses to form pure silicon from sand.

Because ball lightning has not been produced in the laboratory, observational data consist of several thousand accounts from eyewitnesses (see Box 1). An 'average' ball lightning^{1,2} has a diameter somewhere between a golf ball and a large beach ball, and lasts on average 15 seconds (ranging from 2 to 50 seconds) before suddenly fading out or exploding. It can be any colour but is normally white to yellowish and less bright than a 100-watt lightbulb. Ball lightning usually appears out of thin air during thundery weather; its motion is not dictated by the wind and is described as floating in the air not far from the ground. It often bounces when it hits the ground and is influenced by

electric fields. It will scorch wooden objects it strikes, and so is a considerable source of energy.

Previous models of ball lightning have mostly centred on electromagnetically confined plasmas of various kinds, nuclear processes and the chemical burning of gases. In the most recent review of ball lightning³, David Turner observed that "a remarkably consistent picture emerges from the thousands of detailed descriptions which are now available. There is, however, no such consistency in the various hypotheses which have been put forward to explain ball lightning. The only thing most of them share is an ability to explain a few aspects of the phenomenon at the expense of physically impossible requirements in other areas".

The model proposed by Abrahamson and Dinniss² breaks the above trend because it can explain most aspects of ball lightning. The model has three important parts. First, the authors realized that the same chemistry used by the integrated-circuit industry to extract pure silicon from silica-carbon mixtures (SiO_2/C) could be at work in nature, provided that there is one to two times more carbon than silica and a temperature of 3,000 K is reached. Such temperatures are not unusual at the point where lightning strikes. The authors checked the composition of various soils and found that some have the appropriate SiO_2/C ratios. Eureka! So this idea can explain the association of ball



Figure 1 A rare photograph of ball lightning taken by Sankt Gallenkirch in Vorarlberg, Austria, in 1978. This particular example has a whitish centre with a blue surround, and a luminous tail. Such spectacular ball lightning is less frequently seen than the ordinary spherical sort, for which Abrahamson and Dinniss provide a new explanation². There are extremely few photographs of ball lightning, and as yet no video tapes. It is possible that video evidence exists, but that people are unaware of what they have recorded. If anyone reading this has seen, or knows, of such a video tape or photographs, please write or e-mail me.

lightning with thunderstorms, and provide a power source for ball lightning — the chemical energy stored in pure silicon, which is unstable to oxidation at high temperature.

The second part is that the free silicon cools rapidly and condenses into nanoparticles, which can form chains and perhaps even a spherical network of nanostrings, or a 'fluff' ball (my terminology). The authors managed to recreate the conditions of a lightning discharge on soil in the laboratory and produce chains of nanoparticles. But they did not ever see ball lightning. The formation of a ball of nanostrings is the weakest part of this model, but if it can be achieved in future experiments, then we have the remarkable result that most of the properties of ball lightning can be modelled by a fluff ball of silicon. This model can easily explain the floating motions of ball lightning because the silicon networks have very low density, similar to that of air, and it can also explain the range of ball lightning sizes because the initial conditions are quite variable. It is also easy to imagine a small net charge on the ball that would cause it to be influenced by electric fields.

In the third part of their model, the authors calculate the thermal properties of a 30-cm ball of silicon nanostrings. This gives

Box 1: A personal experience

I saw ball lightning during a thunderstorm in the summer of 1960. I was 16 years old. It was about 9 p.m., very dark, and I was sitting with my girlfriend at a picnic table in a pavilion at a public park in upstate New York. The structure was open on three sides and we were sitting with our backs to the closed side. It was raining quite hard. A whitish-yellowish ball, about the size of a tennis ball, appeared on our left, 30 yards away, and its appearance was not directly associated with a

lightning strike. The wind was light. The ball was eight feet off the ground and drifting slowly towards the pavilion. As it entered, it dropped abruptly to the wet wood plank floor, passing within three feet of our heads on the way down. It skittered along the floor with a jerky motion (stick-slip), passed out of the structure on the right, rose to a height of six feet, drifted ten yards further, dropped to the ground and extinguished non-explosively. As it passed my head, I felt no heat. Its acoustic emission I

liken to that of a freshly struck match. As it skittered on the floor it displayed elastic properties (a physicist would call them resonant vibrating modes). Its luminosity was such that it was not blinding. I estimate it was like staring at a less than 10-watt light bulb. The whole encounter lasted for about 15 seconds. I remember it vividly even today, as all eyewitnesses do, because it was so extraordinary. Not until ten years later, at a seminar on ball lightning, did I realize what I had witnessed.

G.K.H.



100 YEARS AGO

The old East Anglian proverb, "As blue as woad," occurs to one visiting the Woad Mill described by Mr. [Francis] Darwin in *Nature*, in 1896 (vol. v. p. 36) as evidence that woad once yielded a blue dye. As a natural sequence one wonders what sort of blue it was and how it was obtained. A somewhat extended series of inquiries amongst those engaged in the woad industry, amongst those who have written on woad, and amongst botanical, archaeological and chemical friends, failed for a long time to elicit the desired information. Curious as it may appear, an appeal to botanical and chemical works, to dictionaries and encyclopaedias was equally unsuccessful. The last-named were pretty uniform in their statements about woad, in that it "was formerly used for dyeing blue, but is now superseded by indigo." Many of the books give an account of the woad-vat in which the manufactured woad is used with bran and lime as a ferment to change the insoluble indigo-blue onto the soluble indigo-white; but they give no clue as to how woad may be used as a blue dye alone. It has been said that the blueness of woad was more or less a myth, and even if it ever possessed this quality it has long since been lost by continued cultivation.

From *Nature* 1 February 1900.

50 YEARS AGO

Eighty years ago, Jevons, then professor of logic at Owens College (now the University of Manchester), built a machine which could perform logical inference by mechanical means. Other similar machines have been built since then. With the present interest in electrical and electronic computing machines, it seemed worth while to construct a logical machine using modern electrical methods, at the same time basing it on the present-day logical technique of truth tables rather more explicitly than had been done by Jevons... It is not to be expected that a machine of this small size will be able to solve logical problems which could not be done with pencil and paper, but it is hoped that this machine may prove to be of value in the teaching of symbolic logic, and that it will stimulate the interest of students in what otherwise tends to become a rather dull subject, and impress on them the mechanical nature of logical operations.

From *Nature* 4 February 1950.

simple explanations for the three most important aspects of ball lightning — lifetime, luminosity and extinction. The lifetime is related to the central starting temperature of the ball after its formation by rapid cooling and condensation, and just before its reheating by oxidation. A lower starting temperature leads to longer lifetimes, calculated to be 2 to 30 seconds. Light emission can last this long because the burn rate is limited by the slow diffusion of oxygen through the developing oxide layer on the surface to the unoxidized silicon underneath.

Abrahamson and Dinniss estimate a luminosity of 1.2 to 14 watts for the silicon ball over the visible range, and the range of blackbody temperatures in the ball's interior can explain most of the colour variation. Their model predicts that heating above a given starting temperature will lead to melting and an explosive end, whereas below a certain starting temperature the ball will completely oxidize before melting and just fade away. Finally, the model predicts that for lower starting temperatures, the ball will become visible only over the latter part of its

lifetime, so the appearance of the ball will not be directly associated with the lightning strike, as is usually observed.

The attractiveness of this model is that it offers a rationale for the duration (very important), delayed time of appearance after lightning strike, luminosity, size, motion and extinction of ball lightning, all of which fit in with my personal experience (Box 1). Ball lightning is such an enigma that new ideas that bear on the problem, such as we find here, are badly needed. Such ideas must then be tested until a definitive result narrows the search. Happily, many of the physical processes in this model are experimentally accessible. We can look forward to observations that will prove or disprove these ideas, which are an unusual but welcome development in research on ball lightning.

Graham K. Hubler is at the Naval Research Laboratory, Code 6370, Washington DC 20375, USA.

e-mail: hubler@ccs.nrl.navy.mil

1. Arago, F. *Meteorological Essays* (transl. Sabine, R. A.) (Brown and Longmans, London, 1855).
2. Abrahamson, J. & Dinniss, J. *Nature* **403**, 519–521 (2000).
3. Turner, D. J. *Phys. Rep.* **293**, 1–60 (1998).

Animal behaviour

Survey flights in honeybees

Thomas Collett

Foraging honeybees may range as far as 10 km from their hive to reach a foraging site and must then find their way home. Before a bee begins its foraging career in earnest, it performs orientation flights that seem to be designed to help it learn landmarks that can guide subsequent returns to the hive. Detailed analysis of these special-purpose flights is helping to clarify the strategies that bees use for learning and navigation. Segments of orientation flights in which the insect is near to the hive can be recorded on videotape — this phase of the flight has been closely studied in ground-nesting wasps when they emerge from their nest holes¹. Later phases, when the bee is far from the hive, are much harder to monitor. Capaldi and her colleagues describe on page 537 of this issue² the first examples of a bee's path during the later phases of the flight. They have used harmonic radar³ to follow the bee from when it leaves the immediate vicinity of the hive to its return.

Individual bees are fitted with a small antenna incorporating a transponder, which, when activated by a radar pulse, emits the first harmonic of the radar signal. The bee can then be picked out from surrounding clutter over a range of 700 m. However, the technique has limitations. It can be used only over open ground, otherwise the bee is masked by vegetation. Positional fixes are

given no more frequently than once every 3 seconds, and even so some fixes are missing. Height is not recorded. Nonetheless, radar tracking is a great advance over earlier methods of investigating long-range navigation. Bees can be tracked by eye for at best 30 m, so earlier studies were limited to measuring journey times and the bearings at which a departing bee vanished from view.

Capaldi *et al.*² find that a typical orientation flight starts with a relatively straight outward path from the hive. After flying between about 10 and 300 m, the bee loops round and returns directly home along a route that is often close to the outward one. Bees make a variable number of these flights (with a mean of about six) before beginning to collect food, with later flights tending to be longer and faster than earlier ones. As individuals have not yet been tracked over multiple flights, it is not known whether a sequence of flights is limited to a narrow sector around the hive. Nor is it known whether orienting bees choose their own flight direction or are directed by the dances of experienced foragers. The relationship between orientation flights and subsequent foraging behaviour will be fascinating to explore.

What do these results reveal about navigational strategies? An important finding is that the longer an orientation flight, the faster the bee flies. This correlation between

Condensed-matter physics

Electrons in the looking glass

Eric Heller

distance and speed suggests that bees fly higher on longer orientation flights. Bees adjust their flight speed to keep images moving at a constant speed across the retina⁴. And to maintain a constant image speed, they will need to fly faster if they gain height. A strategy of increasing height with distance from the hive has interesting consequences. It means that bees learn small features of the landscape close to the ground when they are near the hive, and larger features when they are further away. Their geographical knowledge would be fine-grained close to home and coarse-grained further away, making it ideal for homing in on the hive. Another significant characteristic of orientation flights is the simple hairpin shape of the flight paths. On the straight homeward leg, the bee faces mostly towards the hive, so any views captured en route would be especially helpful in guiding subsequent journeys home. Straight flights towards the hive also allow bees to associate the views they learn with the action of flying along a particular homeward compass bearing.

A bee does learn something about its way home, even from a single orientation flight, as shown first by Becker⁵ and later by Capaldi and Dyer⁶. In the latter study, bees were caught after their first orientation flight and moved to one of several release sites. On release, the bees flew directly towards the hive from those release sites that gave an unobstructed view of the landscape around the hive. Enough landmark information had been acquired during the orientation flight to guide a short homeward trip. The landscape away from the hive is presumably unfamiliar on the orientation flight itself, so the bee's homeward leg is likely to be steered by vector information rather than by landmarks. On the outward leg, the bee records the net distance and direction flown from the hive and then reverses this vector for its return. Straight paths in and out allow the bees to minimize errors in dead reckoning.

That the prime function of these orientation flights is to ensure a safe journey home is so far only a plausible assumption. We can expect this and other issues to be resolved as radar tracking reveals more of the story. An intriguing observation of Capaldi *et al.*² is that bees seem to land during some orientation flights. Why bees land, and whether this behaviour reflects a continuum between orientation and foraging flights, rather than discrete categories, is still uncertain.

Thomas Collett is in the School of Biological Sciences, Biology Building, University of Sussex, Brighton BN1 9QG, UK.

e-mail: t.s.collett@sussex.ac.uk

1. Zeil, J. *J. Comp. Physiol. A* **172**, 189–205 (1993).
2. Capaldi, E. A. *et al. Nature* **403**, 537–540 (2000).
3. Riley, J. R. *et al. Nature* **379**, 29–30 (1996).
4. Srinivasan, M. V., Zhang, S. W., Lehrer, M. & Collett, T. S. *J. Exp. Biol.* **199**, 237–244 (1996).
5. Becker, L. Z. *Vergl. Physiol.* **41**, 1–25 (1958).
6. Capaldi, E. A. & Dyer, F. C. *J. Exp. Biol.* **202**, 1655–1666 (1999).

The scanning tunnelling microscope (STM) is a marvel to scientists and the interested public. Twenty years ago, who would have thought that we would 'see' individual atoms so directly? New science has flowed from this technology like water from a fire hose. After atoms, electrons were next to be imaged, although they are too light to stay put and be 'photographed'. At best the images are blurred, but not just any old blur: the electrons are seen moving in atomic and molecular orbitals, in textbook fashion.

Electrons next made a surprising appearance on the condensed-matter physicist's turf, in the form of conduction electrons freely moving around on the surface of metals such as copper^{1,2}. Not only that, but at low temperatures they showed off their true quantum nature by creating dramatic interference patterns that can only happen when waves come together, adding crest to crest (constructive interference), or crest to trough (destructive interference). The STM images that emerged from Donald Eigler's laboratory are now well known, earning the cover-page 'triple crown': the covers of *Nature*, *Science*

and *Physics Today*^{1–3}. Not surprisingly, the latest STM images from Eigler's group, reported on page 512 of this issue⁴, can again be found on the cover.

It is easier to understand what is going on in these images if we use an acoustic analogy. Suppose you have a mobile omnidirectional speaker, which you place at some point in a room. The speaker plays a sine wave of a given frequency, while you measure power radiated into the room. You would notice a variation in power output by changing either the frequency of the sound or the position of the speaker. The reason is simple: sound reflected off walls and furniture comes back and combines at the speaker to make a returning wave with a given amplitude and phase. This wave makes the speaker do more or less work as it oscillates in and out, depending on the relative phase of the speaker and the returning sound. If you keep the frequency fixed, you could make a plot of power output against speaker position; it would have a definite wavy appearance.

If you have followed this acoustic example so far, you are almost there. Now, replace

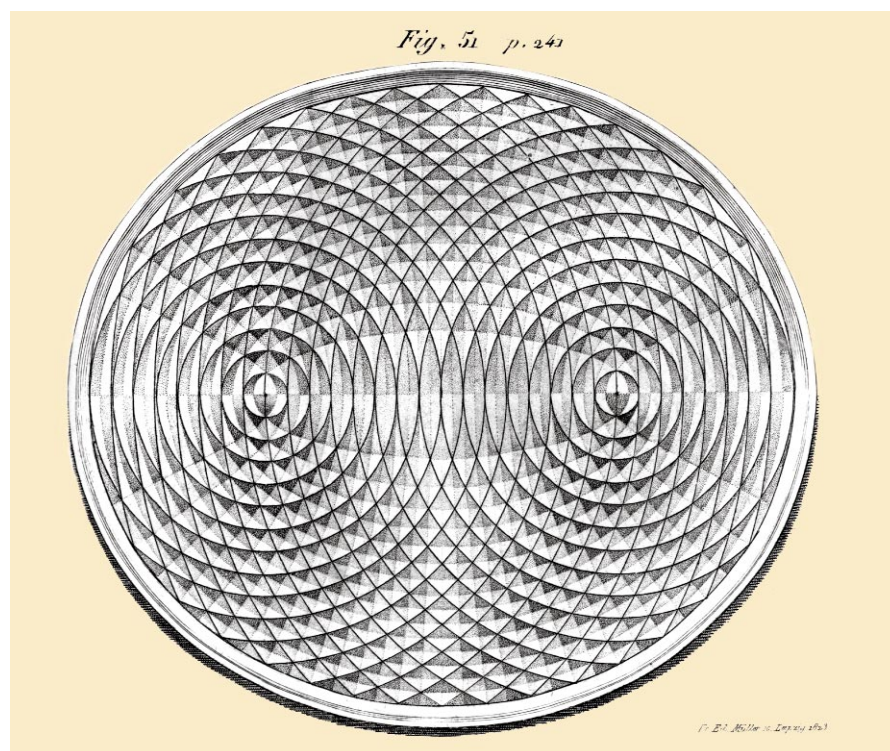


Figure 1 Waves in an elliptical dish of mercury. This drawing by the Weber brothers⁸ in 1825 may be one of the most remarkable hand-drawn scientific illustrations ever. It shows what happens when you put droplets of liquid mercury into one focus of an elliptical dish filled with mercury. The experiment clearly reveals the other focus. In their experiment, Manoharan *et al.*⁴ create an elliptical quantum corral (see cover) in which the signature of an atom at one focus of the ellipse is clearly sensed at the other 'empty' focus.

sound waves with single-electron quantum waves moving over a flat surface of copper. The speaker is the STM tip held just above the surface and supplying electrons; the frequency of the sound is the energy (and corresponding wavelength, controlled by the voltage applied to the tip) of the electrons; and the walls and furniture are edges or other defects, such as iron atoms set in a particular pattern. Finally, power output in the acoustic case becomes current flow in the STM. The famous STM images are the power map of the region scanned by the tip.

The analogy goes further still — real walls and furniture can absorb sound, and windows may be open, so some of the outgoing sound energy is either absorbed or escapes, making the returning signal weaker. In a typical room the sound bounces only a couple of times before it is negligible. That happens with electrons too: in the classic 'quantum corral' arrangement of, say, 60 iron atoms in a circle, the iron walls reflect only about 25% of the wave back into the cavity⁵.

We can now begin to understand the latest experiments by Eigler's group: Manoharan *et al.*⁴ have created a tiny laboratory for remote sensing, in the form of an elliptical quantum corral a few tens of ångströms across. An ellipse is a remarkable geometric object that perfectly focuses either waves or rays starting at one focus onto the other focus (Fig. 1). Manoharan *et al.* put a cobalt atom at one focus and bombarded it with electrons from an STM tip placed at the other focus. Cobalt has what is called a Kondo resonance at low temperatures. This arises because cobalt has a magnetic moment, which at low temperatures causes the 'Fermi sea' of conduction electrons around it (if it is embedded in a metal such as copper) to become oppositely polarized, thereby screening the magnetic moment and producing a many-body singlet state. This Kondo resonance is observed as a rise in resistance at very low temperatures, as the Kondo clouds form, making the magnetic impurities more effective at scattering conduction electrons.

It was Eigler's original dream to see a Kondo resonance up close and personal. The first STM evidence for a Kondo resonance was seen by Madhavan *et al.*⁶, a group led by M. F. Crommie, who did the original quantum corral work with Eigler and Lutz. (Li *et al.*⁷ saw STM surface Kondo resonances around the same time.) The experiments in an elliptical corral reported by Manoharan *et al.*⁴ add a remarkable twist. The Kondo resonance, which manifests itself as a blip in the conductance if the tip is held just above the cobalt atom, has the same (but attenuated) blip if the tip is at the empty focus of the elliptical corral. This raises several fascinating issues. If a Kondo blip is seen at the remote location, does that mean the electronic structure is perturbed there too? Is the electronic structure really being projected to a

remote place, as the authors claim? Or, like a light bulb at one focus of a mirrored elliptical room, do we just see a light bulb at the other focus (as would happen), only to find out that it's not really there when we try to touch it?

We may get some insight by returning to the acoustic analogy. Suppose at one focus of an elliptical room there is a drum that resonates near some frequency, simulating a Kondo resonance. With the speaker on top of the drum, we find a blip in power supplied near the resonance frequency. If we then put the speaker at the other focus, what will happen as we sweep through that same frequency? This is left as an exercise for the reader, as are any conclusions to

be drawn regarding remote projection. ■

Eric Heller is in the Departments of Physics and Chemistry, Lyman Laboratory, Harvard University, 17 Oxford Street, Cambridge, Massachusetts 02138, USA.

e-mail: heller@physics.harvard.edu

1. Crommie, M. F., Lutz, C. P. & Eigler, D. M. *Nature* **363**, 524–527 (1993).
2. Crommie, M. F., Lutz, C. P. & Eigler, D. M. *Science* **262**, 218–220 (1993).
3. *Physics Today* November (1993).
4. Manoharan, H. C., Lutz, C. P. & Eigler, D. M. *Nature* **403**, 512–515 (2000).
5. Heller, E. J., Crommie, M. F., Lutz, C. P. & Eigler, D. M. *Nature* **369**, 464–466 (1994).
6. Madhavan, V. *et al.* *Science* **280**, 567–569 (1998).
7. Li, J., Schneider, W.-D., Berndt, R. & Delley, B. *Phys. Rev. Lett.* **80**, 2893–2896 (1998).
8. Weber, E. H. *Wellenlehre: Auf Experimente Gegrundet oder über die Wellen Tropfbarer Flüssigkeiten mit Anwendung auf die Schall- und Lichtwellen* (Bey Gerhard Fleischer, Leipzig, 1825).

Cancer

Gene expression in diagnosis

Anton Berns

With advances in high-density DNA microarray technology, it has become possible to screen large numbers of genes to see whether or not they are active under various conditions. This is gene-expression profiling, and there has been an expectation that it will revolutionize cancer diagnosis (Box 1)^{1,2}. The thinking is that tumour behaviour is dictated by the expression of thousands of genes, and that micro-array analysis should allow that behaviour and the clinical consequences to be predicted. This rationale is sound enough, but until now it has not been substantiated by experiment.

On page 503 of this issue³, Alizadeh *et al.* deliver such substantiation. The particular cancer they have looked at is diffuse large B-cell lymphoma (DLBCL), a disease that takes in a clinically and morphologically varied group of tumours that affect the lymph system and blood. The authors carried out gene-expression profiling with a

'Lymphochip', a microarray carrying 18,000 clones of complementary DNA designed to monitor genes involved in normal and abnormal lymphocyte development.

Using clustering analysis, Alizadeh *et al.* could separate DLBCL into two categories, which had marked differences in overall survival of the patients concerned. The gene-expression signatures of these subgroups corresponded to distinct stages in the differentiation of B cells, the type of lymphocyte that makes antibodies.

Diffuse large B-cell lymphoma is the most common subtype of non-Hodgkin's lymphoma. With current treatments, long-term survival can be achieved in only 40% of patients. There are no reliable indicators — morphological, clinical, immunohistochemical or genetic — that can be used to recognize subclasses of DLBCL and point to a differential therapeutic approach to patients⁴.

Expression profiling has already shown its usefulness in identifying genes with high

Box 1: Gene-expression profiling with microarrays

Imagine a 1-cm² chessboard. Instead of 64 squares, it has thousands, each containing DNA from a specific gene. This is a DNA microarray. The activity of each gene on the microarray can be compared in two populations of cells (A and B).

When a gene is expressed it makes a transcript, and the whole population of these products from a cell can be

tagged with a fluorescent dye (say, red for the A cells, green for the B cells). The microarray is bathed in a mixture of the red and green transcripts. Those that originate from a specific gene will bind to that gene on the microarray, turning red, green or somewhere in between, depending on the relative numbers of transcripts in the two cell types.

So the microarray provides

a snapshot of gene activity for thousands of genes. Data from many experiments can be compared and genes that have consistent patterns of activity can be grouped or clustered. In this way, genes that characterize a particular cell state, such as malignancy, can be identified — so providing new information about the biology of the cell state.

Mark Patterson

or low expression levels in specific cell types under defined conditions — for instance, when being stimulated with growth factors or treated with drugs, or when the cell's degree of attachment to the extracellular matrix varies (this last characteristic may determine tumour spread)^{2,5–7}. More recently, reports on tumour classification have also begun to emerge. Acute leukaemias can effectively be divided into the lymphoblastic and myeloblastic forms by expression profiling⁸. But in these studies, no multigene-expression signature was found that correlated with a new leukaemia subgroup, or with clinical outcome, in the relatively small group of tumours examined.

In the case of Alizadeh and colleagues' analysis³ of DLBCL, the situation is different. Hierarchical clustering of the gene-expression data divided DLBCLs into two groups: one had the signature of B cells from the germinal centres (the B-cell factories in lymph nodes); the other had the signature of activated B cells. The outlook for patients who had tumours with the activated-B-like signature was much worse — 16 out of 21 died, compared with 6 out of 19 patients with the germinal-centre B-cell signature. Importantly, the predictive value was independent of the standard clinical parameters of prognosis, the International Prognostic Indicator.

That is far from the end of the story, of course. As the authors point out, most of the patients in the 'favourable prognosis' group that die do so within the first two years of diagnosis, whereas some of the patients in the 'poor prognosis' group were still alive after five years. The question is whether there is a 'hidden signature' that, if found, would enable early identification of these subgroups. For the moment, we just cannot say. Testing of more tumours, and using larger or different DNA microarrays, might be needed to resolve this question. In addition, some prognostic indicators might escape detection by expression profiling as they are qualitative, rather than quantitative. That is, genetic (allelic) differences might mean that some genes escape expression screening. They could encode proteins with a different activity or stability that affects tumour progression or response to treatment. In this respect, monitoring of single-nucleotide polymorphisms (SNPs — individual differences at a single base pair that mark a particular genetic variation in the population) would constitute an appealing complementary approach to screening⁹.

When more expression signatures of larger tumour sets become available, it will become clear how this approach will improve monitoring of the stages in which tumours grow and spread, and therefore prognosis. The expectations are high. Furthermore, the better definition of patient groups, made possible by expression profiling, is of obvious importance for assessing the efficacy

of various treatments. Patients will benefit directly from the tailoring of therapies to specific subclasses of tumours.

Finally, gene-expression profiling can be used to identify the genes and pathways that really matter for the tumorigenic process, thereby revealing new targets for therapy. The studies discussed here, and others, show that it is indeed feasible to identify such pathways. For example, high expression of the homeobox gene *HOXA9*, previously identified as a potent oncogene, was correlated with the failure to induce remission in a small group of patients suffering from acute myeloblastic leukaemia⁶. The DLBCL study revealed the increased expression of a series of genes involved in the inhibition of apoptosis (programmed cell death); this, too, might be a therapeutic avenue to explore.

A word of caution. Prognoses based on gene-expression signatures, and — in the future — SNP analysis, are likely to have their limitations. These methods will probably provide quite reliable predictions of the patient responses to initial therapies. The

question then is how much 'noise' will remain as a result of tumour heterogeneity and genetic instability. Both are hallmarks of most malignancies and can lead to recurrent disease, as cancer specialists and their patients know only too well. Time will tell.

For the moment we are witnessing spectacular developments in tumour diagnosis. These developments are going hand in hand with new treatments targeted to defined regulatory pathways that are frequently deregulated in cancer. So although caution is in order, so too is optimism.

Anton Berns is at the Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands.

e-mail: tberns@nki.nl

1. Alizadeh, A. A. *et al.* *J. Clin. Immunol.* **18**, 373–379 (1998).
2. Perou, C. M. *et al.* *Proc. Natl Acad. Sci. USA* **96**, 9212–9217 (1999).
3. Alizadeh, A. A. *et al.* *Nature* **403**, 503–511 (2000).
4. Harris, N. L. *et al.* *J. Clin. Oncol.* **17**, 3835–3849 (1999).
5. Iyer, V. R. *et al.* *Science* **283**, 83–87 (1999).
6. Marton, M. J. *et al.* *Nature Med.* **4**, 1293–1301 (1998).
7. Alon, U. *et al.* *Proc. Natl Acad. Sci. USA* **96**, 6745–6750 (1999).
8. Golub, T. R. *et al.* *Science* **286**, 531–537 (1999).
9. Cargill, M. *et al.* *Nature Genet.* **22**, 231–238 (1999).

Plant ecology

Alien invaders

Peter D. Moore

There are plenty of examples of the ecological mayhem that can result from the human introduction of species into new areas. Stories of goats on Pacific islands, rabbits in Australia and rats in New Zealand are all familiar, as are the world's plant pests, from the aquatic weeds of tropical waterways to the nitrogen-fixing shrubs of Hawaii. The ecological repercussions of plant invasions are often far-reaching, and a fuller understanding of the complexities of ecosystem relationships may provide methods for the control of invasive species. These principles are illustrated by papers in the journals *Conservation Biology* and *Journal of Applied Ecology*. The first deals with the impact of invasive shrubs on songbird breeding success in North America¹; the second with the potential control of unwanted grasses on golf courses in Britain².

Studies of invasive plants often concentrate on the effects of the newcomer on the existing plant community, but its influence may be felt throughout the ecosystem. Schmidt and Whelan¹ have analysed the consequences of the invasion of native North American woodlands by exotic shrubs — a honeysuckle, *Lonicera maaackii*, and buckthorn, *Rhamnus catharticus*. Their particular concern was whether these shrubs are used for nesting by woodland songbirds and, if so, whether nest predation in these species differs from that found in

native shrubs and trees. Both of these invasive shrubs are now common in the eastern part of North America, and both proved to be suitable for and attractive to nesting American robins (*Turdus migratorius*; Fig. 1)



Figure 1 The American robin (*Turdus migratorius*), painted by J. J. Audubon. Schmidt and Whelan¹ found that the robins' nests were more vulnerable to predation when sited in the invading honeysuckle *Lonicera maaackii* than in native trees and shrubs.

and wood thrushes (*Hylocichla mustelina*).

Robins suffered significantly higher nest mortality (mainly as a consequence of the activities of larger mammals, such as raccoons) in *Lonicera* than in native trees and shrubs. Their nests in the *Lonicera* were generally closer to the ground, and overall data relating to nest height above ground showed a negative correlation with nest failure. The difference in architecture and the lack of spines on honeysuckle could also have affected nest predation by mammals; data from native hawthorn (*Crataegus*) species showed no mammal predation at all. A particular cause for concern is the increasing use of *Lonicera* by robins for nesting, despite the higher risks associated with the honeysuckle sites. Over the six-year study, nesting in *Lonicera* increased from 5% of the population to 30%. Perhaps its early leaf production in spring proves deceptively attractive.

Wood thrushes also experienced higher nest mortality in *Lonicera*, but their use of this shrub does not appear to be increasing. Schmidt and Whelan suggest that robins are preventing any such shift by competing effectively with the wood thrushes for *Lonicera* sites. In this respect they may be doing the wood thrushes a big favour. A widely documented decline in songbird populations in North America is giving cause for concern³, but its cause is still unclear. Any factor contributing to such a decline, including reduced nest success in the presence of invasive plant species, deserves close attention.

Introduced species of plants, including these shrubs, are most successful in disturbed, fragmented habitats, where the competitive capacity of many native species may be reduced. Some ecosystems (such as arable fields) are deliberately maintained in a disturbed and simplified state, and these are most susceptible to weed invasion. Another example of such a simplified ecosystem is the putting green of a golf course, where management efforts are aimed at maintaining a close grass sward consisting of relatively few selected grass species. In this situation, invasive species of grasses may also find opportunity for entry. Such is the case with the annual meadow grass or bluegrass (*Poa annua*). The low canopy height of the turf and the presence of small patches of bare soil provide ideal microsites for *Poa* invasion. Once established, this annual grass competes for space with the preferred perennials, such as creeping bent grass (*Agrostis stolonifera*), and then leaves bigger bare patches when it dies at the end of its short life. Its susceptibility to drought and fungal disease also makes this grass less than welcome on a green.

The ability of *Poa* to flower and set seed, even when being cut at a height of 5 mm, provides a constant supply of seed for reinvasion. So, once established, *Poa* is difficult to eliminate, and prevention of initial invasion

may present the best option for its control. Gange, Lindsay and Ellis² analysed all 18 greens on three golf courses and found a negative relationship between the abundance of *Poa* and that of arbuscular mycorrhizal fungi (fungi closely associated with certain plant roots, forming a mutually beneficial relationship). When these fungi were present in profusion, *Agrostis* was also found to perform better, particularly with respect to its root growth. Either the mycorrhizal fungi are enhancing the competitive ability of *Agrostis*, or they are directly harmful to *Poa*. The authors tried a manipulative experiment in which they inoculated a green with mycorrhizal fungi and demonstrated that the root growth of *Agrostis* was stimulated and its abundance increased, giving it an improved chance of resisting *Poa* invasion. They propose, therefore, that the use of this fungal inoculation technique would protect greens against invasion by *Poa* without recourse to chemical or other treatments.

These two studies combine to demon-

strate the complexity of interactions within an ecosystem being invaded by a new species. But this complexity can also be exploited as a means of strengthening an ecosystem's resistance to invasion. Habitat fragmentation, which seems to be an inevitable consequence of modern land use, leaves ecosystems wide open to the establishment of the exotic plants we carry round the world with us. The outcome of invasion by these exotics may be unexpected and will often prove difficult to predict. But a sound knowledge of those ecosystem features that make them resistant to invasion is a priority for conservation biologists. ■

Peter D. Moore is in the Division of Life Sciences at King's College London, Franklin-Wilkins Building, 150 Stamford Street, London SE1 8WA, UK.
e-mail: peter.moore@kcl.ac.uk

1. Schmidt, K. A. & Whelan, C. J. *Cons. Biol.* **13**, 1502–1506 (1999).
2. Gange, A. C., Lindsay, D. E. & Ellis, L. S. *J. Appl. Ecol.* **36**, 909–919 (1999).
3. James, F. C., McCulloch, C. E. & Wiedenfeld, D. A. *Ecology* **77**, 13–27 (1996).

Hydrology

The sting in a fractal tail

Colin P. Stark and Marc Stieglitz

The accident at Chernobyl in April 1986 caused the dispersal of radionuclides, principally of iodine and caesium, across much of Western Europe¹. Contaminated rain fell as far west as the British Isles. During this time, hillwalkers in places such as the Welsh uplands will have looked with greater suspicion than usual at approaching storm clouds. They may also have wondered

how much contamination would enter the soil and groundwater, and how long it would linger — and with good reason. Unfortunately, many aspects of the storage and flushing of groundwater and its contaminants remain poorly understood.

A more innocuous compound that rains on Welsh hikers is sodium chloride, which originates from sea spray blown into the



Figure 1 Plynlimon in mid-Wales, the site of Kirchner and colleagues' work² on catchment flushing. They find that a non-fractal input of sea-salt chlorine in rainfall is converted into a fractal output in stream water — a result that brings into question some basic assumptions about the transit times of water and contaminants in catchment areas.

uplands by onshore winds. This benign contamination has been put to ingenious use by Kirchner *et al.* (page 524 of this issue²), who have examined the storage and flushing of solutes in Plynlimon (Fig. 1), a catchment in mid-Wales. Their aim was to use sea-salt chloride to probe the flushing rates of contaminants from rainfall to streams. They convincingly establish that this watershed exhibits a very clear fractal property: that the apparent travel times of unreactive solutes have a broad-tailed (power law) distribution. This means that some dissolved material will take a long time to be flushed out of the watershed. The general inference is that contaminated catchments may remain contaminated longer than we would expect³.

Rainwater that is not lost through evaporation will eventually find its way to the stream that drains a watershed. But although rain showers or storms are reflected more or less immediately in increased stream discharge, variability in the chemistry of rainfall appears only as very subtle variability in stream chemistry. This has been attributed to the fact that, whereas the pressure of new rainwater (hydraulic head) can induce movement of older water into the stream as an immediate response to rain, the older water exhibits the integrated chemical signal of rainwater received over long periods of time.

Chloride is relatively non-reactive in groundwater, so it can be used as a conservative tracer for waters as it passes from rain to groundwater to the stream. By analysing time series of chloride concentration in rainfall and stream flow, Kirchner *et al.*² are therefore able to make some inferences about the time that a tracer takes to work its way through the system. In particular, they can deduce the distribution of apparent travel times of water in the catchment.

A typical simplifying trick in stochastic hydrology is to treat the catchment and its groundwater reservoir as a black box that acts as a simple spectral filter. Rainfall is fed into the system as a time series, and the catchment black box filters this into a time series that simulates stream flow. The same idea can also be applied to stream chemistry if the solute is unreactive.

This useful simplification lies at the heart of the time-series analysis of the Plynlimon study. The input (rainfall) chloride signal to this watershed has an approximately flat spectrum. The output chloride signal has a power-law spectrum, where the spectral energy is inversely proportional to frequency, which is diagnostic of what is technically termed a $1/f$ noise, or fractal signal. The black-box filter therefore appears to convert a flat input spectrum into a fractal output spectrum. This is one of the strong appeals of the Plynlimon case: it is very rare to find good examples in nature of systems which so clearly turn a non-fractal input into a fractal output.

If the fractal filtering property of conservative tracers can be demonstrated in catchments elsewhere, we may need to rethink our basic assumptions about the transit times of water and contaminants in shallow groundwater. A common black-box model assumes a simple, well-mixed reservoir, in which a single spike input of solute leads to an exponentially decaying yield of solute output. The equivalent distribution of travel times in this kind of reservoir has an exponential shape. The flushing of contaminants in such a system is more efficient than in a fractal reservoir — the yield of unreactive solutes is higher immediately after contamination, and diminishes relatively quickly. The inference to be made from the Plynlimon catchment is that something different occurs: a lower initial yield of tracer, and a more sustained yield over many subsequent years.

But the story is not quite as simple as that. A fundamental limitation of the spectral black-box approach is that it does not describe the response of the system to a point injection of contaminant, as would occur for example if a drum of benzene were spilt on a hillside. Nor does it directly describe the flux of contaminant over time during a single contamination episode. Rather, the spectral approach aggregates many such episodes, and assumes a reasonably spatially uniform input of contaminant. So how does the

inferred fractal property of the catchment pertain to a single, point injection of contaminant? This remains to be seen.

On the other hand, this work will have wide application, despite its spatial imprecision. It is likely that its main use will be as a kind of baseline constraint for catchment-scale simulations of chemical transport. Modellers of solute flow will need to bear this fractal filtering in mind, regardless of whether the solute is reactive or not. Applications will no doubt emerge in models of the flushing of sulphates from acid rain, transport of dissolved organic carbon and nitrogen, and residence times of helium isotopes (helium-3).

As for radioisotopes such as caesium-137, deposited in catchments following events such as the Chernobyl disaster, the complexities of fractal filtering are perhaps moot: caesium remains essentially fixed in the top few centimetres of the soil, bound to clays and other substrates. Another worry, perhaps, for those mud-soaked hikers in the Welsh hills.

Colin P. Stark and Marc Stieglitz are at the Lamont-Doherty Earth Observatory of Columbia University, Palisades, New York 10964, USA.

e-mail: cstark@ldeo.columbia.edu

1. Barnaby, F. *Ambio* **15**, 332–334 (1986).
2. Kirchner, J. W., Feng, X. & Neal, C. *Nature* **403**, 524–527 (2000).
3. Schnoor, J. L. *Environmental Modeling: Fate and Transport of Pollutants in Water, Air, and Soil* (Wiley, New York, 1996).

Huntington's disease

A predictor of pathology

John C. Rothwell

Like woodworm spreading through a ship's timbers, most slowly progressive diseases begin long before clinical symptoms appear. Although modern techniques can often detect preclinical changes — and, potentially, can identify how and where a disease begins — there is one main problem. Patients rarely complain of being ill before they experience symptoms, so it can be tricky to find a suitable population to test. Reporting in this issue, however, Smith and colleagues have neatly circumvented these difficulties in their study of the genetically inherited movement disorder Huntington's disease (Smith, M. A., Brandt, J. & Shadmehr, R. *Nature* **403**, 544–549; 2000).

Smith *et al.* have identified asymptomatic carriers of the Huntington's disease gene, and predict that these people will develop this movement disorder within the next five to ten years. The authors tested subjects for subtle deficits in the control of skilled arm movements, and, extraordinarily, found changes in their ability to point accurately at

targets up to seven years before the predicted onset of clinical symptoms. Given that the pathology so early in the disease is probably

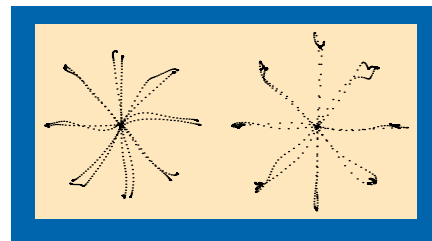


Figure 1 Examples of arm movements made from a central starting position to eight peripheral targets in a healthy subject (left) and an asymptomatic carrier of the Huntington's disease gene (right). The tracks of two attempts at each movement are superimposed. Smith *et al.* have shown that asymptomatic carriers correct their trajectories much more than normal as they approach the targets, leading to greater jerkiness in the final stages of the movements. This is seen as much as seven years before the predicted onset of the disease.

limited to a relatively small population of neurons within the basal ganglia (an area of the brain beneath the cerebral cortex that is involved in controlling movement), this observation gives an unexpected insight into the function of these neuronal structures in normal movement.

George Huntington of Pomeroy, Ohio, described his eponymous disease in 1872. The disease usually begins at between 30 and 40 years of age, with symptoms of clumsiness and subtle involuntary movements that gradually become more florid and are accompanied by a slow intellectual decline. It is an autosomal dominant inherited motor disorder, caused by repeated copies of the trinucleotide sequence CAG in the *IT15* gene on the short arm of chromosome 4. Pathologically, the initial neural damage is limited to the medium spiny neurons of the caudate and putamen within the basal ganglia. The pathology later spreads to involve the cerebral cortex and other structures.

In the years before the affected gene was identified, symptoms of the disease often had to be quite severe before a definite diagnosis could be made. At this late stage, it was never clear to what extent the movement disorder reflected the pathology within the basal ganglia or in other structures. Smith *et al.* have now circumvented this problem by studying carriers of the Huntington's disease gene before their symptoms begin.

The authors asked people carrying the affected gene and normal subjects to move a two-jointed object with their arm, from a central position to one of eight targets on a surrounding circle (Fig. 1). When asymptomatic gene carriers carried out the task, the initial trajectory of their arm movements seemed to be relatively normal. But as they approached the target, almost all movements failed to stop efficiently and smoothly. Smith *et al.* quantified how smooth each movement was by calculating the mean squared jerk along its course (jerk is the time derivative of acceleration). They found that the amount of extra jerk in the asymptomatic gene carriers was inversely proportional to the expected time of onset of clinical symptoms, as calculated from the length of the CAG trinucleotide repeat and the age of their parents when the carriers were born. In effect, the smoothness of the terminal error correction could be used to predict the approximate time at which symptoms of disease would begin.

Normal subjects make small corrections to their initial arm trajectories as they approach the target, meaning that jerkiness occurs towards the end of a movement. The authors found that asymptomatic gene carriers seem to have no problem in calculating the initial trajectory of the movement. Instead, these people appear to have trouble implementing the appropriate corrections that are needed towards the end point. So, in

a second experiment, Smith and colleagues deliberately perturbed the initial trajectory of the arm movements. Again, the corrective movements of asymptomatic gene carriers were disturbed far more dramatically than expected. This result was quite different from that seen in a separate group of patients with disease of the cerebellum, who also had clumsy movements. When these patients were analysed in the pointing task, the initial trajectory was more irregular than usual but the reaction to the external perturbation was normal.

Because the pathology in asymptomatic carriers of the Huntington's disease gene is probably limited to the basal ganglia, the implication is that these neuronal structures might be involved in implementing error corrections during arm movement. Indeed, Smith *et al.* propose that pathology of the basal ganglia may affect how the cortex processes sensory signals relating to the magnitude of the error. However, as in all studies of chronic disease processes, there is an alternative possibility. The observed movement deficit may not itself be caused by the primary pathology in the basal ganglia. As the disease develops, other parts of the brain may increase their function to

compensate. So any deficit might actually reflect a secondary side effect of compensation, rather than being a result of the initial pathology. To decide between the two possibilities we need further evidence in healthy people for activity in the basal ganglia during this type of task.

As the human genome is slowly revealed, the approach of identifying subtle subclinical deficits in asymptomatic gene carriers can be used in more and more neurological diseases. It may allow us to develop simple predictive tests of symptom onset, and to study the physiological effects of pathological processes that are confined to a relatively small part of the central nervous system. But a classic ethical dilemma remains. Until we have a treatment for these diseases — and Huntington's disease, with its relentless downward clinical course, is currently untreatable — we, and the people who participated in this experiment, are no better than the blind prophet Tiresias, who commented: "It is but sorrow to be wise when wisdom profits not".

John C. Rothwell is at the MRC HMBU, Institute of Neurology, Queen Square, London WC1N 3BG, UK.

e-mail: j.rothwell@ion.ucl.ac.uk

Molecular dynamics

Optical control of reactions

Stuart A. Rice

Active control of the internal motions of a molecule, by use of a tailored optical field (such as a laser pulse), can be achieved by exploiting a variety of interference effects associated with the quantum mechanics of molecular motion. For example, the population of a specified vibrational state, which may be a precursor to the end-product of a chemical reaction, can be controlled using a timed sequence of laser pulses, as was first suggested 15 years ago^{1,2}. The theories behind several other control methods were worked out at about the same time^{3–5} and are still being actively developed^{6–10}.

A few years ago the first experimental confirmations of the proposed control methods began to appear^{11,12}, and the art of optical control of molecular dynamics is now advancing rapidly¹³. An example of the state of the art in pulse-timing control of molecular dynamics is provided by a report in *Chemical Physics Letters* by Frohnmeyer *et al.*¹⁴, in which they directly map the atomic separation of sodium atoms in a sodium dimer by using intense femtosecond laser pulses.

The fundamental time scale of chemical reactions is set by the internal vibrations of

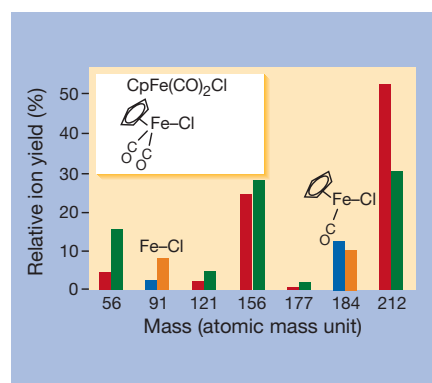


Figure 1 Selection of bond-breaking in a polyatomic molecule using femtosecond laser pulses. Starting from the parent molecule ($\text{CpFe}(\text{CO})_2\text{Cl}$), two different bond-breaking reactions are chosen that lead to chemically different final products (FeCl^+ and CpFeCOCl^+). In the experiment, the ratio of competing products ($\text{CpFeCOCl}^+ / \text{FeCl}^+$) is maximized (blue) or minimized (orange) by an algorithm that sends a feedback signal to control the laser pulses. In each case, the relative yield of CpFeCOCl^+ and FeCl^+ is significantly different. The abundances of the parent molecule (red) and any other fragments (green) are not included in the feedback signal.

the reacting molecules, which are incredibly fast — typical vibration periods are 10–100 femtoseconds. (One femtosecond is 10^{-15} seconds, which is to a second what a second is to 32 million years.) It was not possible to observe such reactions in real time until the development of femtosecond laser pulses. In one type of experiment, two laser pulses are used to monitor the molecular motions. The first 'pump' pulse excites the molecule to a higher energy state with a particular type of motion, and the subsequent 'probe' pulse examines what is happening.

In Frohnmeyer *et al.*'s experiment¹⁴, the probe pulse is also used to ionize the excited molecule, producing an electron. The amount of energy carried away by this electron depends on the separation of the atoms in the molecule. By varying the delay between the two pulses, Frohnmeyer *et al.* are able to record the motion of the sodium atoms in the excited molecule. Moreover, by increasing the intensity of the probe pulse they observe systematic changes in the bonding between the sodium atoms, which is one way of using femtosecond pulses to control chemical reactions.

One of the most interesting and rapidly advancing uses of active control of molecular dynamics is for product selection in a branching chemical reaction (Fig. 1). Most chemical reactions produce more than one product, and the art of synthetic chemistry is devoted to improving the yield of the desired product. Some desired products can only be produced with difficulty and with poor yield using conventional methods. If the reaction could always be channelled towards that product, the efficiency of synthesis would be greatly enhanced. Advances in femtosecond lasers in the past five years, especially in the technology for shaping laser pulses with respect to spectral content, phase distribution and temporal shape, have made such control possible for simple reactions.

So far, control of product selection using timed pulses has already been demonstrated for branching between photo-fragmentation and photo-ionization of the sodium dimer¹¹. A different control method has been used to select between the competing processes of photo-fragmentation and photo-ionization of hydrogen iodide¹². In this case, interference between optical fields used to excite two independent transitions between the same initial and final states was exploited to control product selection. In another experiment, different excited states generated by the photo-fragmentation of the sodium dimer were selected by generating coherent (in-phase) oscillations of population between a small set of molecular levels¹⁵. Finally, control of bond breaking in a polyatomic molecule using an optimally phase-controlled pulse has been demonstrated for the competing formation of FeCl^+ and CpFeCOCl^+ from the photo-

fragmentation of $\text{CpFe(CO)}_2\text{Cl}$ (Fig. 1)¹³.

Although our understanding of the requirements for active control of molecular dynamics is well founded, there remain many fundamental questions that must be addressed. For example, is there a fundamental limit to the control of many-body quantum dynamical processes? How does the efficiency of a control method depend on the number of possible motions of the controlled molecule? And how sensitive is the control field to source fluctuations, other experimental imperfections and uncertainties in the molecular potential energy? To what extent is control of molecular dynamics possible when dissipation of energy must be accounted for, as when the reaction is carried out in solution? Moreover, is it possible to control collision processes, so as to affect the choice of products in a bimolecular reaction¹⁶?

The existing theories and experiments only hint at what will be possible in the future. For some time to come, applications using control of molecular dynamics are likely to be most important as tools for learning more about molecular dynamics and for testing and developing better theories. But as laser technology improves, and our understanding of complex molecules advances, practical applications will be developed. In the short term, the most likely applications will be in optoelectronics — we have already seen optical control of the microscopic state of a quantum dot¹⁷ and of a fast semiconductor switch based on interference between optical excitation pathways^{18,19}. What next for control of molecular dynamics? ■

Stuart A. Rice is at the James Franck Institute, University of Chicago, 5640 South Ellis Avenue, Chicago, Illinois 60637, USA.
e-mail: sarice@rainbow.uchicago.edu

1. Tannor, D. J. & Rice, S. A. *J. Chem. Phys.* **83**, 5013–5018 (1985).
2. Tannor, D. J., Kosloff, R. & Rice, S. A. *J. Chem. Phys.* **85**, 5805–5820 (1986).
3. Shapiro, M. & Brumer, P. *J. Chem. Phys.* **84**, 4103–4104 (1986).
4. Peirce, A. P., Dahleh, M. A. & Rabitz, H. *Phys. Rev. A* **37**, 4950–4964 (1988).
5. Kosloff, R., Rice, S. A., Gaspard, P., Tersigni, S. & Tannor, D. J. *J. Chem. Phys.* **139**, 201–220 (1989).
6. Shapiro, M., Chen, Z. & Brumer, P. *J. Chem. Phys.* **217**, 325–340 (1997).
7. Kobrak, M. N. & Rice, S. A. *Phys. Rev. A* **57**, 2885–2894 (1998).
8. Kobrak, M. N. & Rice, S. A. *J. Chem. Phys.* **109**, 1–10 (1998).
9. Charon, E., Giusti-Suzor, A. & Mies, F. H. *Phys. Rev. Lett.* **75**, 2815–2818 (1995).
10. Manz, J. & Paramanov, G. K. *J. Phys. Chem.* **97**, 12625–12633 (1993).
11. Baumert, T., Elbing, J. & Gerber, G. *Adv. Chem. Phys.* **101**, 47–77 (1997).
12. Fiss, J. A., Zhu, L. C., Suto, K., He, G. Z. & Gordon, R. J. *J. Chem. Phys.* **233**, 335–345 (1998).
13. Assion, A. *et al. Science* **282**, 919–922 (1998).
14. Frohnmeyer, T., Hofmann, M., Strehle, M. & Baumert, T. *J. Chem. Phys. Lett.* **312**, 447–454 (1999).
15. Shnitman, A., Sofer, I., Golub, I., Yegor, A. & Shapiro, M. *Phys. Rev. Lett.* **76**, 2886–2889 (1996).
16. Abrashkevich, A., Shapiro, M. & Brumer, P. *Phys. Rev. Lett.* **81**, 3789–3792 (1998).
17. Bonadeo, N. H. *et al. Science* **282**, 1473–1476 (1998).
18. Kurizki, G., Shapiro, M. & Brumer, P. *Phys. Rev. B* **39**, 3435–3444 (1989).
19. Dupont, E., Corkum, P. B., Liu, H. C., Buchanan, M. & Wasilewski, Z. R. *Phys. Rev. Lett.* **74**, 3596–3599 (1995).

Daedalus

Perfect dryness

Daedalus once devised a mist-proof plastic for lenses, windows, and so on. He charred its surface with a laser and treated it with fluorine, converting it to graphite fluoride, the most water-repellent substance known. Water drops simply could not adhere. He now plans to extend the idea to wood, leather and similar organics.

Such materials are porous. Water does not merely condense on them; it penetrates, and acts as a medium for biological decay. But how to char their vast internal porosity with a few monolayers of graphite?

Strong sulphuric acid is notorious for charring organic materials. It abstracts the elements of water from their molecules, leaving carbon. Sulphur trioxide, says Daedalus, is in effect dehydrated sulphuric acid. It should extract hydrogen and oxygen from organics even more avidly. Being easily volatile, it could permeate into the most porous material. Dilution with inert gas could limit its charring action to a few monolayers of surface. Diluted fluorine would then convert the graphite to its fluoride. All internal porosities, however complex, would be perfectly waterproofed.

DREADCO chemists are now trying it. They are exposing oars, pairs of boots, plywood shingles, woollen socks, carpets and similar objects to sequences of vapours in a large pressure-vessel. When the process has been perfected, these products will emerge utterly waterproof. Even under high pressure, water will not wet them, let alone enter their pores. But like lesser 'breathable' waterproof fabrics (some of which are also based on fluorocarbon polymers), they will still be freely permeable to air and water vapour.

This elegant process will transform many technologies. The British building industry is dominated by fear of wood-rot. The first waterproofed, unpainted wooden buildings, flaunting their appealing natural grain against the now impotent weather, will arouse amazement and disbelief. So will truly waterproof boots, fabrics that need no washing (and indeed cannot be washed — water, like dirt, simply falls off them), and wooden boats that slip unwetted through the water. Daedalus even hopes to waterproof human skin. A waterproof swimmer would break all Olympic records. The slow renewal of the skin would gradually remove the coating.

David Jones

The Further Inventions of Daedalus (Oxford University Press), 148 past Daedalus columns expanded and illustrated, is now on sale. Special *Nature* offer: m.curtis@nature.com

Obituary

Olivier Kahn (1943–99)

Olivier Kahn, a pioneer and protagonist of molecular-based magnetic materials, and a major influence on contemporary inorganic chemistry, died suddenly in Paris on 8 December of an aneurysm at the early age of 57. He had just returned from a visit to Japan, where he was collaborating with a multidisciplinary group at the Institute of Physical and Chemical Research (RIKEN), and the following day was due to chair a meeting of the French Chemical Society on coordination chemistry. These are just two examples, in his always-crowded diary, of the international reach of his interests and his intellectual roots as an inorganic coordination chemist.

Olivier Kahn played a key part in the transformation of inorganic chemistry, over the past 20 years, from a science largely concerned with individual molecules to one in which greater emphasis is placed on the behaviour of bulk material. He was born and grew up in Paris in a family in which intellectual debate, in the highest tradition of French metropolitan life, was part of the atmosphere (his brother is a noted social and political commentator on French television). Graduating first in his class in chemistry from the Ecole Nationale Supérieure de Chimie, he carried out his doctoral research at the same school. Particularly influential in developing his ideas was a postdoctoral stay in Britain, at the University of East Anglia, in the early 1970s, where he learned about the quantum-mechanical approach to magnetic interactions from the chemist Sid Kettle.

The 1950s and 1960s had witnessed a renaissance in inorganic chemistry, driven by attempts to first synthesize and then rationalize the structures and properties of new compounds, in which metal atoms (usually from the transition elements) were surrounded by organic molecules called ligands that modulate both structure and reactivity. At the time, the dominant theoretical framework for understanding inorganic properties was ligand field theory. This theory necessarily focused attention on single metal atoms and their immediate surroundings (usually organic ligands) — a mindset that became even more entrenched with the explosive flowering of organometallic chemistry.

Cracks were already beginning to appear in this purely molecular view of inorganic chemistry. They were fuelled in part by a resurgence of interest in the classical solid-state chemistry of oxides



Magnetochemist who helped create a new approach to inorganic chemistry

and other continuous lattice compounds, pioneered by J. S. Anderson and others. It was also beginning to be appreciated how important interactions between molecules can be in determining properties of matter in bulk, such as cooperative magnetic and optical phenomena in one- or two-dimensional solids. Into this epoch burst the figure of Olivier Kahn.

Kahn grasped the opportunity presented by his appointment as the youngest full professor at the University of Paris-Sud, Orsay, to build up (and, in the expressive Gallic phrase, 'animate') a new laboratory. He was eager, following his apprenticeship in coordination chemistry and molecular quantum mechanics, to create and articulate a new subject.

Early success came with the design and synthesis of a molecular complex containing one element from the beginning (vanadium) and one from the end (copper) of the first transition series, placed side by side so that orbitals containing the unpaired electrons were strictly orthogonal to one another. This arrangement was predicted by P. W. Anderson, M. Kanamori and J. B. Goodenough to lead to a parallel alignment of the spins, resulting in ferromagnetism (that is, each molecule becomes strongly magnetized, like permanent magnets such as iron). It was a triumph. A ferromagnetic ground state was observed, though not of course one where the ordering was long range —

this was molecular ferromagnetism.

In quick succession, Kahn made further marks on coordination chemistry by two contrasting *démarches*. First, with B. Briat and others, he recast the Anderson–Kanamori–Goodenough model of exchange interactions between localized magnetic moments into a form accessible to chemists. Second, he extended his synthetic success with ferromagnetic dimers to infinite chains of transition-metal ions coupled through organic bridges to make ferrimagnetic arrays, in which the spins are aligned antiparallel, but unequal contributions still produce magnetization. In this way, Kahn connected the molecular world with the physicists' paradigm of infinite one-dimensional magnetic order.

—Magnetic chains may be infinite in one sense, but are not themselves true bulk magnets because the chains must interact with each other to give three-dimensional order. It was then that Kahn showed, by a beautiful piece of supramolecular chemical tweaking, how the chains in his one-dimensional ferrimagnets could be arranged so as to give ferromagnetic interactions between the chains. Other lattice topologies followed, including one with three-dimensionally interlocking rings of metal atoms joined together by organic ligands. This one Kahn compared, with a characteristic flight of fancy, to a necklace that he had bought for his wife.

Throughout his career, first in Orsay and later Bordeaux, where he established a large and flourishing laboratory in the 1990s, Kahn led his young team by his own lively example: exclamations of delight at the latest findings of a graduate student echoed through the corridors. When a colleague chided him gently for spending little time at home, his reply was "but we are always together on Sunday afternoons".

For Olivier Kahn, finding beautiful new patterns in nature — whether strung together by necklaces of atoms in a crystal lattice or in his recent theories of molecular bi-stability as a basis for information storage — was an all-consuming passion, which he shared ebulliently with everyone he met. For his friends across the globe, as well as for the molecular sciences he espoused so eloquently, the hole left by his passing will be hard to fill. Not to see that eager figure, bolt upright in the centre of the front row in the lecture theatre, spectacles gleaming, poised to ask the first question, makes our science duller as well as sadder. **Peter Day**

Peter Day is at the Davy Faraday Research Laboratory, The Royal Institution of Great Britain, 21 Albemarle Street, London W1X 4BS, UK.
e-mail: pday@ri.ac.uk

Longevity record for deep-sea invertebrate

The growth rate of a marine tubeworm is tailored to different environments.

Communities around deep-sea hydrothermal vents experience rapidly fluctuating and ephemeral environments that stimulate their biological growth and development at rates far exceeding those of other deep-sea communities^{1–4}. In contrast, the slow, steady release of reduced compounds from hydrocarbon seeps provides a much more constant environment, although, surprisingly, these support communities of species that are phylogenetically and physiologically quite similar to those at the vents^{5–8}. Here we show that one of the creatures living around hydrocarbon seeps on the Louisiana continental slope, the ecosystem-structuring vestimentiferan tubeworm, *Lamellibrachia* sp., requires between 170 and 250 years to grow to a length of two metres. This amazingly slow growth is in marked contrast with that of its vent relatives, who are among the fastest-growing invertebrates on the planet⁴. It also makes *Lamellibrachia* sp. the most long-lived non-colonial marine invertebrate known.

To estimate growth rates and longevity in *Lamellibrachia* sp., we established study sites at 27° 47' N, 91° 30' 24" W and 27° 44' 46" N, 91° 13' 18" W, at depths of 540 m and 560 m, respectively. Groups of vestimentiferan tubeworm tubes in 22 different aggregations were marked in 1994, 1995 and 1997 with a blue chitin stain (acid blue 148) using an *in situ* device⁹ manipulated by the *Johnson Sea Link* manned submersible. Previously stained individuals were collected in 1995, 1997 and 1998, and new tube growth was measured as the length of the



Figure 1 Patch of vestimentiferan tubes stained with blue chitin stain. New tube growth appears as the unstained anterior portion of the tube. Even within a single stained patch (~25 cm in diameter), growth varies tremendously.

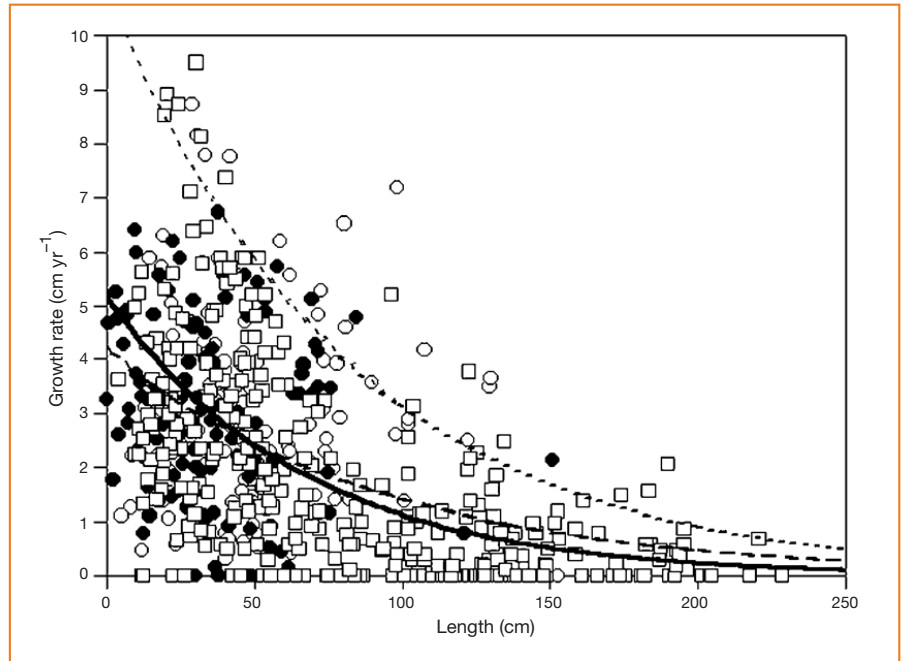


Figure 2 Relation between growth rate and standardized tube length of *Lamellibrachia* sp. Stained individuals collected in 1995 (filled circles), 1997 (open circles) and 1998 (squares). Negative exponential model fit to the 1998 data (solid line), the entire data set (dashed line) and the upper bound of the 1998 data (dotted line). The whole data set consisted of all *Lamellibrachia* sp. stained and collected between 1994 and 1998, including those with estimated standardized lengths (calculated from the relationship between anterior tube diameter and tube length) and those with an interval of one, two and three years between staining and collection. One data subset ('1998') consisted of animals stained in 1997 and collected in 1998, for which standardized length did not require estimation; this subset covered the full size range of animals collected in this study and contained the fewest potentially confounding factors. The second data subset ('upper bound of 1998') consisted of individuals displaying the highest 10% of the growth rates in each 20-cm size class in the 1998 data; this subset provides the theoretical maximum growth rate.

unstained segment of tube between the stain mark and the anterior end (Fig. 1), and then standardized to a yearly growth rate. Tube length, determined by following the curvature of the tube with a piece of twine, was standardized by measuring from the anterior edge of the stain mark to the posterior point at which the tube tapered to 2 mm in outer diameter.

Collections of small, entirely stained aggregations show the anterior end to be the only region of new tube deposition above the point of attachment of the tube to the substrate. *Lamellibrachia* sp. can also grow downwards from below the point of attachment of its tube to the substrate; however, as all our measurements were made above the point of attachment, growth below this point did not affect the age estimates reported here.

We used nonlinear regression to fit a variety of growth models to all the *Lamellibrachia* sp. growth-rate data and to two subsets of these data (Fig. 2). A simple negative exponential model of the form $dL/dt = ae^{-bL}$, where a and b are parameters and L is the standardized animal length,

consistently provided the best and most robust fit. Age was then estimated for each individual, i , by integrating the negative exponential model and solving $t = (1/ab)e^{bL}$ over the definite integral (zero, standardized length of i).

Our results indicate that the ages of animals in the upper range of sizes collected for this study (over 2 m long) are in excess of 170–250 years. Even if it were assumed that individuals could maintain their maximum growth rate throughout their lives, the larger animals would still be over 80 years old. However, our data indicate that individuals do not grow at their maximal rate throughout their lives: growth occurs episodically¹⁰, and similarly sized individuals in a single stained patch within an aggregation show virtually the full range of growth rates observed for that size class (Fig. 1). Because of constraints upon sampling with the submersible, our analysis does not include data from the largest individuals at these study sites. Our lifespan estimates of 170–250 years for *Lamellibrachia* sp. are therefore conservative.

The deep sea is a mainly nutrient-poor

environment, where vertical habitat structure is often non-existent above the sea floor. In contrast, aggregations of long-lived *Lamellibrachia* sp. at hydrocarbon seeps in the Gulf of Mexico are able to provide a stable, long-term source of habitat, settlement substrate and nutrition for a host of endemic and transient fauna. The evolution of a very long lifespan in this slow-growing, cold-seep vestimentiferan species is particularly intriguing when considered alongside the remarkably rapid growth of vestimentiferans adapted to living around ephemeral hydrothermal vents.

Derk C. Bergquist, Frederick M. Williams, Charles R. Fisher

Department of Biology,
Pennsylvania State University,
University Park,
Pennsylvania 16802, USA

1. Grassle, J. F. *Adv. Mar. Biol.* **23**, 301–362 (1986).
2. Tunnicliffe, V. *Oceanogr. Mar. Biol. Annu. Rev.* **29**, 319–407 (1991).
3. Sarrazin, J., Robigou, V., Juniper, S. K. & Delaney, J. R. *Mar. Ecol. Prog. Ser.* **153**, 5–24 (1997).
4. Lutz, R. A. *et al.* *Nature* **371**, 663–664 (1994).
5. Kennicutt, M. C. II, Brooks, J. M., Bidigare, R. R. & Denoux, G. J. *Deep-Sea Res.* **35**, 1639–1651 (1988).
6. Paull, C. K. *et al.* *Science* **226**, 965–967 (1984).
7. Anderson, R. K., Scanlon, R. S., Parker, P. L. & Behrens, E. W. *Science* **222**, 619–622 (1983).
8. Hecker, B. *Biol. Soc. Wash. Bull.* **6**, 465–473 (1985).
9. Urcuyo, I. A., Massoth, G. J., MacDonald, I. R. & Fisher, C. R. *Cah. Biol. Mar.* **39**, 267–270 (1998).
10. Fisher, C. R., Urcuyo, I. A., Simpkins, M. A. & Nix, E. *Mar. Ecol.* **18**, 83–94 (1997).

Reproductive biology

Mitochondria and the death of oocytes

In females of many species, over half of the germ-cell (oocyte) population dies by apoptosis before birth¹. For example, germ-cell numbers peak at $5\text{--}7 \times 10^6$ at week 20 of gestation in humans, but drop to less than 1×10^6 in the early neonatal period^{2,3}. Apparent germ-cell wastage occurs on a similar scale in female rodents, falling from 6.4×10^4 at day 17.5 of pregnancy to 1.9×10^4 shortly after birth⁴. Krakauer and Mira⁵ have interpreted this death of germ cells as a developmental solution to the accumulation of mutations in mitochondria, proposing that prenatal oocyte apoptosis effectively removes oocytes carrying mutant mitochondria. Here we test whether mitochondria can actually influence oocyte fate by microinjecting small numbers of mitochondria into mouse oocytes and find that this prevents these cells from undergoing apoptosis. We also show that a common mitochondrial DNA deletion occurs more frequently in unfertilized, as compared with fertilized, human oocytes.

We used oocytes from FVB female mice, because these oocytes undergo inherently high rates of apoptosis *in vitro*⁶. After being denuded of somatic (granulosa) cells, each oocyte was microinjected individually, either with buffer or with about 5×10^3 mitochondria purified from non-apoptotic follicular granulosa cells of female mice that had been primed 46 hours beforehand with a single injection of equine chorionic gonadotropin to promote granulosa-cell viability⁷. After culturing for 24 hours, 70% of the oocytes that had either not been microinjected or had been microinjected with buffer underwent apoptosis^{8,9}, whereas only 36% of the oocytes microinjected with mitochondria initiated apoptosis (Fig. 1). As a single mouse oocyte contains about 1×10^5 mitochondria¹⁰, these findings are

striking, considering that the total mitochondrial pool per microinjected oocyte was only increased by some 5%.

We then used the polymerase chain reaction to assay for a common deletion of 4,977 base pairs in mitochondrial DNA in 72 human primordial follicles isolated from ovarian biopsies of women aged 20–49. Deletions were detected in 14 follicles (R.G.G. *et al.*, unpublished results), although without the correlation with age reported in long-lived somatic cells¹¹. In parallel studies, the frequency of deletions in human embryos was lower than in oocytes, suggesting that superior oocytes had been selected for fertilization (R.G.G. *et al.*, unpublished results).

Although these two pieces of evidence support the hypothesis of Krakauer and Mira⁵, several issues remain unresolved. We agree that one function of prenatal germ-cell loss could be to remove oocytes with defective mitochondrial genomes; however, these oocytes probably represent a very small percentage of the total oocyte pool

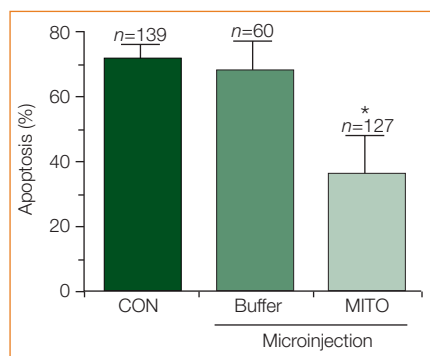


Figure 1 Effect of microinjecting mitochondria on the occurrence of spontaneous apoptosis in murine oocytes incubated *in vitro* for 24 hours. The magnitude of apoptosis was equivalent in oocytes not microinjected (control, CON) and those microinjected with buffer alone. However, microinjection of approximately 5×10^3 purified mitochondria (MITO) suppressed the activation of programmed cell death in cultured oocytes by 50%. The total number of oocytes used in each group is indicated over the respective bar (mean $\pm$ s.e.m., $n = 3$ independent experiments; * $P < 0.05$).

lost before birth, and thus additional functions for prenatal oocyte death need to be considered¹.

The proposed mitochondrial mutations must also be defects carried forward from the previous generation(s), as it is implausible that so many oocytes could be lost prenatally as a result of mitochondrial DNA mutations accumulating between oogenesis and parturition. This may create a paradox, however, because Krakauer and Mira⁵ predict that the organism will arise from an oocyte free of mitochondrial DNA defects. In many animal species, the mitochondria present in all cells of the organism are derived from replication of the original maternal (oocyte-derived) mitochondria, implying that all cells in the newly developing embryo, including the germ line, would possess mitochondria free of DNA mutations.

Krakauer and Mira⁵ also do not account for the continued postnatal loss of oocytes in many species. The oocyte population in human females declines mainly as a result of the degeneration (atresia) of follicles housing each oocyte to about 3×10^5 at puberty, 2.5×10^4 at age 37, and 1×10^3 at menopause¹². A similar situation occurs in postnatal rodent ovaries, although the number of follicles is proportionately smaller¹³.

The events responsible for the initiation of postnatal follicular atresia do not universally depend on the oocyte as the prime factor. Atresia is driven by oocyte apoptosis in the early (primordial, primary, preantral) stages of development^{6,13,14}, but somatic granulosa cells of the follicle are the first to undergo apoptosis during atresia of follicles at later stages of development, including the point of preovulatory selection¹⁴. Krakauer and Mira⁵ do not explain how mutations in the mitochondrial DNA of an oocyte might trigger apoptosis in the surrounding granulosa cells as a means to remove that follicle from the ovulatory pathway, particularly in the light of the fact that gonadotropins prevent atresia of maturing follicles by direct action on the follicular somatic cells¹⁴.

Gloria I. Perez*, Alexander M. Trbovich*, Roger G. Gosden†, Jonathan L. Tilly*

*Vincent Center for Reproductive Biology,
Department of Obstetrics and Gynecology,
Massachusetts General Hospital/Harvard
Medical School, VBK137E-GYN, Boston,
Massachusetts 02114, USA

e-mail: jtilly@partners.org

†Centre for Reproduction, Growth and
Development, University of Leeds, Leeds General
Infirmary, Leeds LS2 9NS, UK

1. Morita, Y. & Tilly, J. L. *Dev. Biol.* **213**, 1–17 (1999).
2. Baker, T. G. *Proc. R. Soc. Lond. B* **158**, 417–433 (1963).
3. Forabosco, A. *et al.* *Anat. Rec.* **231**, 201–208 (1991).
4. Beaumont, H. M. & Mandl, A. M. *Proc. R. Soc. Lond. B* **155**, 557–579 (1961).

5. Krakauer, D. C. & Mira, A. *Nature* **400**, 125–126 (1999).
6. Morita, Y. *et al.* *Mol. Endocrinol.* **13**, 841–850 (1999).
7. Tilly, J. L. *et al.* *Endocrinology* **136**, 232–241 (1995).
8. Perez, G. I. *et al.* *Nature Med.* **3**, 1228–1232 (1997).
9. Perez, G. I. *et al.* *Mol. Hum. Reprod.* **5**, 414–420 (1999).
10. Piko, L. & Matsumoto, L. *Dev. Biol.* **49**, 1–10 (1976).
11. Lee, H. C. *et al.* *Biochim. Biophys. Acta* **1226**, 37–43 (1994).
12. Faddy, M. J. *et al.* *Hum. Reprod.* **7**, 1342–1346 (1992).
13. Perez, G. I. *et al.* *Nature Genet.* **21**, 200–203 (1999).
14. Tilly, J. L. & Robles, R. in *Molecular Biology in Reproductive Medicine* (ed. Fauser, B. C. J. M.) 79–101 (Parthenon, New York, 1999).

Krakauer and Mira reply — Many of the issues raised by Perez *et al.* concern molecular mechanism rather than evolutionary function. Thus ‘cause’ does not carry the same meaning in evolution as in physiology. The evolutionary cause or function of sexual reproduction or two-step meiosis can be seen as a means of fending off parasites or selfish genetic elements: the developmental or mechanistic causes of these are different and involve precise chemical and genetic processes. We propose that the evolutionary function of atresia has been to eliminate mildly deleterious mutations from mitochondria, thereby retarding Müller’s ratchet¹. It is not necessary for all of the oocytes undergoing apoptosis in each generation to contain defective mitochondrial genomes, much as individuals do not need to be teeming with parasites or sister-killer mutants to engage in sex, or for cells to undergo meiosis. Once these adaptations arise, organisms are often committed to their implementation².

We suggest that the long-term evolutionary persistence of atresia requires that germ cells with healthy mitochondria represent statistically superior competitors in the ovary and are thus more likely to escape cell death to seed successive generations. The result of this competition is to eliminate the worst cells (when present) and retain a sufficiently large population of cells to produce the maximum number of offspring. This is rather like a game of musical chairs, where the worst competitors are eliminated early on and there is only one winner among many equally matched contestants. Thus many egg cells compete for survival factors through metabolism, but only a few will be successful. The mechanism remains unclear, although oocyte elimination must involve initiation of apoptosis by some somatically derived signal.

A key part of our theory is that competition for survival factors be promoted among germ cells carrying a small number of mitochondrial genomes. Cells with large numbers of defective mitochondria can metabolize at almost 100% efficiency, providing at least 10% of wild-type mitochondrial genomes are present³ (the ‘threshold effect’), so selective differences among cells

only become apparent when wild-type genomes are rare in the germ cell. This is the function of the bottleneck — it decreases mitochondria to a level that makes the functional variance among cells selectively detectable (Fig. 1b of ref. 1).

The results of Perez *et al.* agree with our hypothesis and match our comparative data, showing that the number of mitochondria present during the bottleneck can predict the level of atresia. Both experimental and comparative data indicate that small numbers of healthy mitochondria can inhibit atresia, which in our view is evidence that mitochondria can indirectly influence oocyte fate. Mitochondrial genome-deletion experiments confirm the importance of genome quality for mitochondrial persistence.

Perez *et al.* suggest that atresia in the next generation is redundant, as the founder population of oocyte mitochondria have previously experienced selection. What is important selectively is the stage in development when mitochondria are sequestered into the oocytes and the reproductive lifespan of the mother. Assuming that all mitochondria in the oocyte at reproductive maturity are a subset of those present in pre-atretic fetal development (not the case in all species), then mutations could accumulate (through division or the effects of environmental mutagens) from the time of atresia up until the formation of the next-generation fetus. This can be between 13 and 40 years in humans. There is good evidence that ontogenetic mutations accumulate in mitochondria to reconstitute polymorphism⁴, which might contribute by a similar mechanism towards the ongoing postnatal atresia discussed by Perez *et al.* In the absence of a germline bottleneck, ontogenetic ageing of mitochondria should proceed more rapidly.

As Perez *et al.* point out, many more cells are eliminated during early atresia than might plausibly contain defective mitochondrial genomes. This is surprising, but selection may incur some continuous or facultative response proportional to the mutational load per generation. As we noted earlier, however, this is not a feature of mammalian meiosis or sex that is obligate for development. The adaptively salient outcome of atresia is that it retains sufficient eggs for the female’s reproductive lifetime. This might simply be a matter of insurance: evolution cannot always discover the globally optimal solution.

We have offered a purely functional theory of atresia, one that has the merit of addressing an evolutionary puzzle, Müller’s ratchet in mitochondria. We believe that evolution (function) and molecular biology (mechanism) must work together in this area.

David C. Krakauer*, Alex Mira†

*Institute for Advanced Study, Princeton, New Jersey 08540, USA

†Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA

1. Krakauer, D. C. & Mira, A. *Nature* **400**, 125–126 (1999).
2. Maynard Smith, J. & Szathmáry, E. *The Major Transitions in Evolution* (W. H. Freeman, Oxford, 1995).
3. Chomyn, A. *et al.* *Proc. Natl Acad. Sci. USA* **89**, 4221–4225 (1993).
4. Corral-Debrinsky, M. *et al.* *Nature Genet.* **2**, 324–329 (1992).

Embryogenesis

Demethylation of the zygotic paternal genome

In mammals, both parental genomes undergo dramatic epigenetic changes after fertilization to form the diploid somatic genome. Here we show that the paternal genome in the mouse is significantly and actively demethylated within 6–8 hours of fertilization, before the onset of DNA replication, whereas the maternal genome is demethylated after several cleavage divisions. This active demethylation of the paternal genome may be associated with epigenetic remodelling of sperm chromatin, in order to establish parent-specific developmental programmes during early embryogenesis.

Previous molecular analysis^{1–3} of digested genomic DNA from mature germ cells indicated that ovulated mouse oocytes are globally undermethylated compared with the sperm genome. A gradual genome-wide demethylation seems to occur during pre-implantation development, leading to indistinguishable alleles at most gene loci, but not at those that are imprinted^{1,2}.

Intermediate amounts of methylation could be caused by a combination of undermethylated maternal and methylated paternal DNA³, although they could be the result of highly dynamic and sometimes opposing demethylation or *de novo* methylation processes in parental genomes⁴. To investigate this, we analysed the global methylation of the paternal and maternal genomes in mouse pre-implantation embryos by using immunofluorescence against 5-methylcytosine (MeC). Because most MeC is found in various repeat-DNA families^{4,5}, antibody staining mainly reflects the density of MeC in interspersed repetitive sequences.

At 3 hours (Fig. 1a) and 6 hours (Fig. 1b) after fertilization, both the paternal and maternal chromosomes stained equally intensely with anti-MeC antibody. Because the ovulated oocytes are globally undermethylated, this finding suggests that there is rapid *de novo* methylation of egg chromatin shortly after fertilization¹. After 8 hours, the paternal pronuclei were decondensed and so were much larger than the maternal pronuclei. Unexpectedly, all

paternal pronuclei analysed after 8 hours showed very little methylation (Fig. 1c).

To prevent DNA replication, one-cell embryos were collected after 6 hours and cultured for a further 10 hours in the presence of aphidicolin (at $2 \mu\text{g ml}^{-1}$)⁶. The unreplicated paternal pronuclei also became MeC-negative (Fig. 1d). This demethylation may be associated with transient hyperacetylation of histone H4 (ref. 7), because both replication and transcription are initiated earlier in the male pronucleus⁸, which is less condensed.

At first metaphase, we observed two differentially methylated and spatially separated chromosome sets⁹ (Fig. 1e). To exclude the possibility that differential MeC staining was due to changes in the accessibility of paternal DNA, we stained mouse embryos with anti-DNA antibody¹⁰: male and female pronuclei produced equally intense

immunofluorescence (Fig. 1f–j). Two-cell embryos in phases G1 and G2 of the cell cycle were prepared at 22 and 32 hours. Most interphase nuclei displayed highly localized MeC staining (Fig. 1k,l), reflecting the compartmentalization of the two genomes, which may bring about the differential epigenetic reprogramming of the two genomes.

The amount of global methylation of the maternal genome was largely maintained from the early pronuclear to the two-cell stage, but four-cell embryos 45 hours after fertilization had a much lower MeC density over the maternal half of the nucleus (Fig. 1m). Interphase nuclei of 16- and 32-cell embryos had equivalently low methylation of paternal and maternal DNA (data not shown). Thus, in contrast to the very rapid and active demethylation of the paternal pronucleus, gradual demethylation of the

maternal genome occurred passively during the second and third cleavage stages by a replication-dependent mechanism⁹, which may involve the loss of maintenance methylase activity. The second polar body remained methylated throughout pre-implantation development.

Our results provide a dramatic demonstration of the loss of DNA modification after fertilization. Active zygotic demethylation of the paternal genome has important implications for the understanding of genomic imprinting, X-chromosome inactivation, mammalian cloning and *in vitro* fertilization.

Wolfgang Mayer*†, Alain Niveleau‡, Jörn Walter*, Reinald Fundele*, Thomas Haaf*

*Max Planck Institute of Molecular Genetics, Ihnestr. 73, 14195 Berlin, Germany
e-mail: haaf@molgen.mpg.de

‡Molecular and Structural Virology Unit, Université J. Fourier de Grenoble, 38706 La Tronche, France

†Deceased.

1. Howlett, S. K. & Reik, W. *Development* **113**, 119–127 (1991).
2. Olek, A. & Walter, J. *Nature Genet.* **17**, 275–276 (1998).
3. Monk, M. *et al. Development* **99**, 371–382 (1987).
4. Yoder, J. A., Walsh, C. P. & Bestor, T. H. *Trends Genet.* **13**, 335–340 (1997).
5. Bird, A. P. *Nature* **321**, 209–213 (1986).
6. Howlett, S. K. *Roux's Arch. Dev. Biol.* **195**, 499–505 (1986).
7. Adenot, P. G. *et al. Development* **124**, 4615–4625 (1997).
8. Bouniol-Baly, C. *et al. Exp. Cell Res.* **236**, 201–211 (1997).
9. Rougier, N. *et al. Genes Dev.* **12**, 2108–2113 (1998).
10. Scheer, U. *Eur. J. Cell Biol.* **43**, 358–371 (1987).

Correction

Fluid 'rope trick' investigated

L. Mahadevan, W. S. Ryu, A. D. T. Samuel
Nature **392**, 140 (1998)

We wish to amend a small mistake in our calculations, although this does not affect the basic idea in our paper, particularly by comparison with experiment. As the longitudinal viscous stress σ in the filament scaled as $\sim \mu U r / R^2$ varies linearly across the cross-section, the integrated stress resultant $\int \sigma dA$ vanishes. However, the net bending torque due to this viscous stress does not vanish and scales as $\int \sigma r dA \sim \mu U r^4 / R^2$. The force per unit volume on the fluid due to centripetal and Coriolis accelerations scales as $f \sim \rho \Omega^2 R$, so that the bending torque on the whirling filament in the vicinity of the coil scales as $f r^2 R^2 \sim \rho \Omega^2 r^2 R^3$. Torque balance, together with the ancillary continuity relations, leads to a scaling law for the coiling frequency

$$\Omega \sim Q^{4/3} r^{-10/3} \nu^{-1/3} \quad (1)$$

which is slightly different from the result given in our paper, where an erroneous argument confuses the transverse and longitudinal timescales in the filament. Equation (1) can be derived directly using an analogy to the coiling of an elastic rope by simply replacing the elastic bending modulus $E r^4$ in ref. 1 with the 'viscous bending modulus' $\mu^4 U / R$. A reconsideration of the experimental results leads to data collapse with a power law $\Omega / Q^{1/33} \sim r^{-3.45 \pm 0.10}$, in agreement with our argument.

1. Mahadevan, L. & Keller, J. B. *Proc. R. Soc. Lond. A* **452**, 1679–1694 (1996).

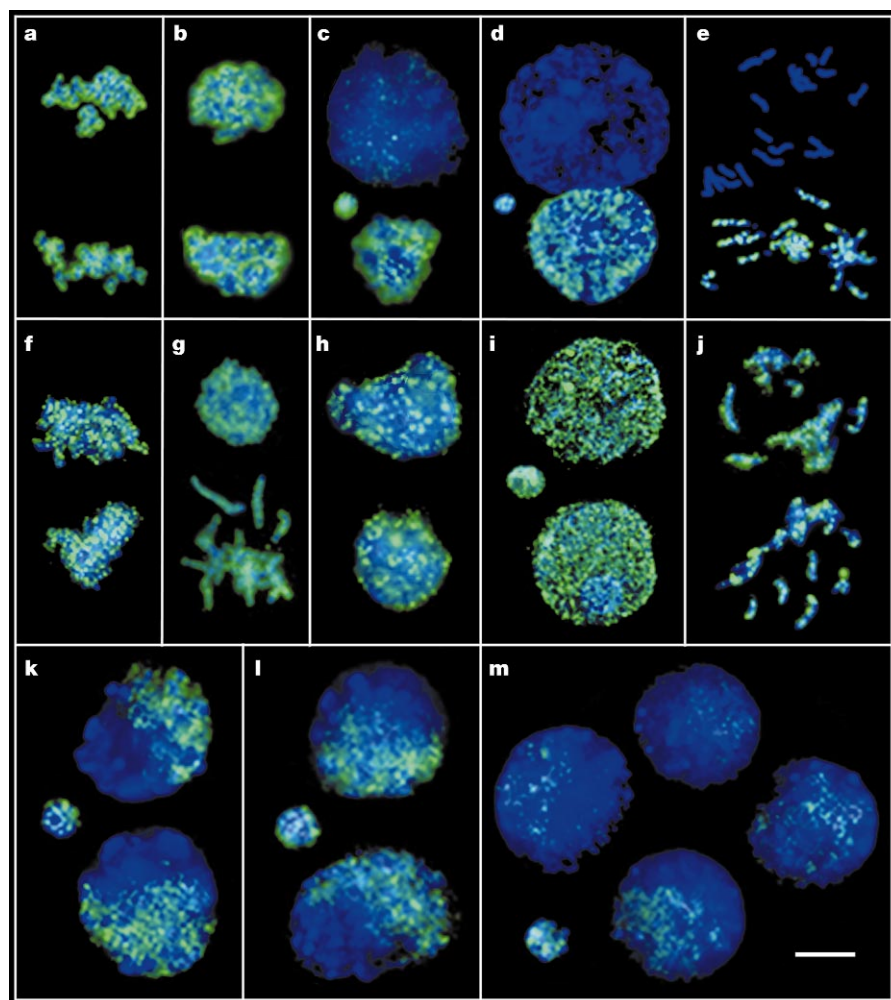


Figure 1 Differential demethylation of parental chromatin in the early mouse embryo. **a–e**, Anti-5-methylcytosine (MeC) immunofluorescence of mouse one-cell embryos. **a**, Zygote 3 h after fertilization with intense MeC labelling of both pronuclei (>10). Numbers in parentheses indicate the number of embryos analysed. **b**, Paternal and maternal pronuclei at 6 h (>10). **c**, Undermethylated paternal pronucleus at 8 h (>20). The smaller female pronucleus remains methylated. **d**, Aphidicolin-treated one-cell embryo displaying demethylation of the male pronucleus (>20). **e**, First metaphase (>5). **f–j**, Controls. Anti-DNA immunofluorescence of one-cell embryos demonstrates that both chromatin sets are accessible to antibody molecules. **f**, 3 h (>5). **g**, 6 h (>5). **h**, 8 h (>5). **i**, Aphidicolin treatment (>5). **j**, First metaphase (2). **k,l**, MeC staining of two-cell embryos at 22 h (>20) (**k**) and 32 h (>20) (**l**) shows that the paternal and maternal compartments have different methylation levels. **m**, Four-cell embryo at 45 h (>10). The MeC-staining intensity of the maternal half of the nucleus is weaker than in two-cell embryos. Scale bar, 10 μm .

Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling

Ash A. Alizadeh^{1,2}, Michael B. Eisen^{2,3,4}, R. Eric Davis⁵, Chi Ma⁵, Izidore S. Lossos⁶, Andreas Rosenwald⁵, Jennifer C. Boldrick¹, Hajeer Sabet⁵, Truc Tran⁵, Xin Yu⁵, John I. Powell⁷, Liming Yang⁷, Gerald E. Marti⁸, Troy Moore⁹, James Hudson Jr⁹, Lisheng Lu¹⁰, David B. Lewis¹⁰, Robert Tibshirani¹¹, Gavin Sherlock⁴, Wing C. Chan¹², Timothy C. Greiner¹², Dennis D. Weisenburger¹², James O. Armitage¹³, Roger Warnke¹⁴, Ronald Levy⁶, Wyndham Wilson¹⁵, Michael R. Grever¹⁶, John C. Byrd¹⁷, David Botstein⁴, Patrick O. Brown^{1,18} & Louis M. Staudt⁵

Departments of ¹Biochemistry, ²Genetics, ³Pathology, ⁴Medicine, ⁵Pediatrics and ¹¹Health Research & Policy and Statistics, and ¹⁸Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California 94305, USA

⁵Metabolism Branch, Division of Clinical Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

⁷Bioinformatics and Molecular Analysis Section, CBEL, CIT, NIH, Bethesda, Maryland 20892, USA

⁸CBER, FDA, Bethesda, Maryland 20892, USA

⁹Research Genetics, Huntsville, Alabama 35801, USA

Departments of ¹²Pathology and Microbiology, and ¹³Internal Medicine, University of Nebraska Medical Center, Omaha, Nebraska 68198, USA

¹⁵Medicine Branch, Division of Clinical Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

¹⁶Johns Hopkins Oncology Center, Johns Hopkins School of Medicine, Baltimore, Maryland 21287, USA

¹⁷Walter Reed Army Medical Center, Washington, DC 20307, USA

²These authors contributed equally to this work

Diffuse large B-cell lymphoma (DLBCL), the most common subtype of non-Hodgkin's lymphoma, is clinically heterogeneous: 40% of patients respond well to current therapy and have prolonged survival, whereas the remainder succumb to the disease. We proposed that this variability in natural history reflects unrecognized molecular heterogeneity in the tumours. Using DNA microarrays, we have conducted a systematic characterization of gene expression in B-cell malignancies. Here we show that there is diversity in gene expression among the tumours of DLBCL patients, apparently reflecting the variation in tumour proliferation rate, host response and differentiation state of the tumour. We identified two molecularly distinct forms of DLBCL which had gene expression patterns indicative of different stages of B-cell differentiation. One type expressed genes characteristic of germinal centre B cells ('germinal centre B-like DLBCL'); the second type expressed genes normally induced during *in vitro* activation of peripheral blood B cells ('activated B-like DLBCL'). Patients with germinal centre B-like DLBCL had a significantly better overall survival than those with activated B-like DLBCL. The molecular classification of tumours on the basis of gene expression can thus identify previously undetected and clinically significant subtypes of cancer.

Despite the variety of clinical, morphological and molecular parameters used to classify human malignancies today, patients receiving the same diagnosis can have markedly different clinical courses and treatment responses. The history of cancer diagnosis has been punctuated by reassortments and subdivisions of diagnostic categories. There is little doubt that our current taxonomy of cancer still lumps together molecularly distinct diseases with distinct clinical phenotypes. Molecular heterogeneity within individual cancer diagnostic categories is already evident in the variable presence of chromosomal translocations, deletions of tumour suppressor genes and numerical chromosomal abnormalities. The classification of human cancer is likely to become increasingly more informative and clinically useful as more detailed molecular analyses of the tumours are conducted.

The classification of human lymphomas has steadily evolved since their initial recognition by Thomas Hodgkin in 1832 (ref. 1). Beginning with the distinction of Hodgkin's disease from other malignant and non-malignant conditions^{2,3}, a variety of lymphoma classifications have been advanced on the basis of both morphologic and molecular parameters⁴. The most recent classification scheme, the Revised European-American Lymphoma (REAL) classification, was introduced to categorize distinct clinical-pathological entities⁵.

However, within this classification system, various morphologic subtypes were unified into groups despite the suspicion that they "include more than one disease entity"⁵.

Diffuse large B-cell lymphoma (DLBCL) is one disease in which attempts to define subgroups on the basis of morphology have largely failed owing to diagnostic discrepancies arising from inter- and intra-observer irreproducibility^{5,6}. Diffuse large B-cell lymphoma is an aggressive malignancy of mature B lymphocytes, with an annual incidence of over 25,000 cases, accounting for roughly 40% of cases of non-Hodgkin's lymphoma. Patients with DLBCL have highly variable clinical courses: although most patients respond initially to chemotherapy, fewer than half of the patients achieve a durable remission^{6,7}. Although a combination of clinical parameters is currently used to assess a patient's risk profile, these prognostic variables are considered to be proxies for the underlying cellular and molecular variation within DLBCL⁸.

An important component of the biology of a malignant cell is inherited from its non-transformed cellular progenitor. Each of the currently recognized categories of B-cell malignancy has been tentatively traced to a particular stage of B-cell differentiation, although the extent to which these malignancies maintain the molecular and physiological properties of normal B-cell subsets is not clear. The rearranged immunoglobulin genes in DLBCL and most other non-Hodgkin's lymphomas bear mutations that are characteristic of somatic hypermutation, an antibody-diversification

⁴Present address: Life Sciences Division, Lawrence Berkeley National Labs and Department of Molecular and Cellular Biology, University of California, Berkeley, California 94720, USA.

mechanism that normally occurs only within the germinal centre of secondary lymphoid organs⁹. This evidence suggests that DLBCL arises either from germinal centre B cells or from B cells at a later stage of differentiation.

Here we examined the extent to which genomic-scale gene expression profiling can further our understanding of B-cell malignancies. We addressed whether we could (1) generate a molecular portrait of distinct types of B cell malignancy; (2) identify distinct types of B-cell malignancy not recognized by the current classification system; and (3) relate each malignancy to normal stages in B-cell development and physiology. We focused particularly on DLBCL to determine whether gene expression profiling could subdivide this clinically heterogeneous diagnostic category into molecularly distinct diseases with more homogeneous clinical behaviours.

Construction of a specialized DNA microarray

Recent technical and analytical advances make it practical to quantitate the expression of thousands of genes in parallel using complementary DNA microarrays¹⁰. This mode of analysis has been used to observe gene expression variation in a variety of human tumours^{11–17}. To apply this method to questions in normal and malignant lymphocyte biology, we designed a specialized microarray—the ‘Lymphochip’—by selecting genes that are preferentially expressed in lymphoid cells and genes with known or suspected roles in processes important in immunology or cancer¹⁸.

Because of the suspected importance of the germinal centre B cell to the genesis of non-Hodgkin’s lymphomas, 12,069 out of the 17,856 cDNA clones on this microarray were chosen from a germinal centre B-cell library¹⁸. An effort was made to include all distinct genes that were initially discovered in this library. We included an additional 2,338 cDNA clones from libraries derived from DLBCL, follicular lymphoma (FL), mantle cell lymphoma and chronic lymphocytic leukaemia (CLL). Finally, we added clones representing a variety of genes that are induced or repressed during B- and T- lymphocyte activation by mitogens or cytokines¹⁹ and a curated set of 3,186 genes of importance to lymphocyte and/or cancer biology. About a quarter of the genes included in this microarray were represented by two or more different cDNA clones, providing internal controls for the reproducibility of gene expression quantitation. See Supplementary Information for the complete annotated list of these cDNAs.

Analysis of gene expression in lymphoid malignancies

We used these microarrays to characterize gene expression patterns in the three most prevalent adult lymphoid malignancies: DLBCL, FL and CLL (Fig. 1). To provide a framework for interpretation of the gene expression in these patient samples, we also profiled gene expression in purified normal lymphocyte subpopulations under a range of activation conditions, in normal human tonsil and lymph node, and in a variety of lymphoma and leukaemia cell lines. Fluorescent cDNA probes, labelled with the Cy5 dye, were prepared from each experimental messenger RNA sample. A reference cDNA probe, labelled with the Cy3 dye, was prepared from a pool of mRNAs isolated from nine different lymphoma cell lines. Each Cy5-labelled experimental cDNA probe was combined with the Cy3-labelled reference probe and the mixture was hybridized to the microarray. The fluorescence ratio was quantified for each gene and reflected the relative abundance of the gene in each experimental mRNA sample compared with the reference mRNA pool. The use of a common reference probe allowed us to treat these fluorescent ratios as measurements of the relative expression level of each gene across all of our experimental samples.

In all, ~1.8-million measurements of gene expression were made in 96 normal and malignant lymphocyte samples using 128 Lymphochip microarrays. Figure 1 provides an overview of the variation in gene expression across these samples. A hierarchical clustering algorithm was used to group genes on the basis of similarity in the

pattern with which their expression varied over all samples²⁰. The same clustering method was used to group tumour and cell samples on the basis of similarities in their expression of these genes. The data are shown in a matrix format, with each row representing all the hybridization results for a single cDNA element of the array, and each column representing the measured expression levels for all genes in a single sample. To visualize the results, the expression level of each gene (relative to its median expression level across all samples) was represented by a colour, with red representing expression greater than the mean, green representing expression less than the mean, and the colour intensity representing the magnitude of the deviation from the mean²⁰.

Distinct clones representing the same gene were typically clustered in adjacent rows in this gene map, indicating that these genes have characteristic and individually distinct patterns of expression and showing that the effects of experimental noise or artefact are negligible. Likewise, where different tumour samples from the same patient were analysed, they were invariably found clustered in immediately adjacent columns. For example, in three cases of FL in which the malignant cells were separated from the normal host cells by magnetic cell sorting, the purified and unpurified samples from the same patient clustered next to each other. Two samples of leukaemic cells from the same CLL patient were obtained 18 months apart, and these samples were more highly correlated in gene expression with each other than with any other patient’s CLL cells. The observed patterns of gene expression thus reflected intrinsic differences between the tumours, rather than variation in handling or experimental artefacts. Moreover, these results show that even within a diagnostic category, each cancer patient has a unique tumour with a characteristic gene expression profile.

Figure 1 paints a complex, but remarkably ordered, picture of the variation in gene expression patterns in lymphoid malignancies, with large sets of genes displaying coordinate expression in related biological samples. Although no information on the identity of the samples was used in the clustering, the algorithm segregated, with few exceptions, the recognized classes of lymphoid malignancies based on global similarities in gene expression patterns. Examination of the coordinately expressed genes in each of the B-cell malignancies and comparison with the normal lymphocyte cell populations yielded considerable insights into the biology of these malignancies. The coloured bars on the right of Fig. 1 indicate clusters of coordinately expressed genes that we operationally defined as gene expression ‘signatures’. A gene expression signature was named by either the cell type in which its component genes were expressed (for example, the ‘T-cell’ signature) or the biological process in which its component genes are known to function (for example, the ‘proliferation’ signature). Thus, the overall gene expression profile of a complex clinical sample such as a DLBCL lymph-node biopsy can be understood, in a first approximation, as a collection of gene expression signatures that reveal different biological features of the sample.

Gene expression patterns and tumour phenotype

One of clearest distinctions between the gene expression patterns of the three B-cell malignancies involved genes that vary in expression with cellular proliferation rates. Both CLLs and FLs were clustered next to resting B-cell samples, which reflects, in part, the fact that both of these malignancies are relatively indolent, with very low proliferation rates. Correspondingly, the genes that define the proliferation signature were not highly expressed in these malignancies (Fig. 2). This gene expression signature included diverse cell-cycle control genes, cell-cycle checkpoint genes, DNA synthesis and replication genes, and the gene Ki67, commonly used to gauge the ‘proliferation index’ of a tumour biopsy, as previously noted¹⁵. In general, the more rapidly proliferating DLBCLs had higher expression of the genes in the proliferation signature. Nonetheless, marked differences in the expression of these genes were evident

between individual DLBCL samples, corresponding to the variability in proliferation index that has been previously observed in DLBCL²¹.

The most prominent distinction between CLL and FL came from genes that are characteristic of germinal centre B cells (Fig. 2). An extensive cluster of genes distinguished germinal centre B cells from both resting blood B cells and *in vitro* activated blood B cells. This is remarkable because the stimuli used to activate the blood B cells were chosen to mimic those known to be important for germinal centre formation: crosslinking of the immunoglobulin receptor and CD40 signalling. However, it has thus far not been possible to mimic exactly the germinal centre phenotype *in vitro*, as determined by the failure of a variety of activation conditions to induce the expression of BCL-6 protein, a highly specific marker for germinal centre B

cells²². The germinal centre B-cell gene expression signature shows that germinal centre B cells represent a distinct stage of B-cell differentiation and not merely one specific form of B-cell activation. Support for this notion comes from the fact that the characteristic gene expression program of germinal centre B cells was maintained in a cultured DLBCL cell line in the absence of the germinal centre microenvironment (Figs 1 and 2).

The observation that FLs show a pattern of ongoing somatic hypermutation of immunoglobulin genes has led to the suggestion that the transformation event leading to FL occurs while the B cell is in the germinal centre microenvironment²³. The gene expression signature of germinal centre B cells was reproduced virtually unchanged in FL, supporting the view that this lymphoma arises from this stage of B-cell differentiation (Fig. 2).

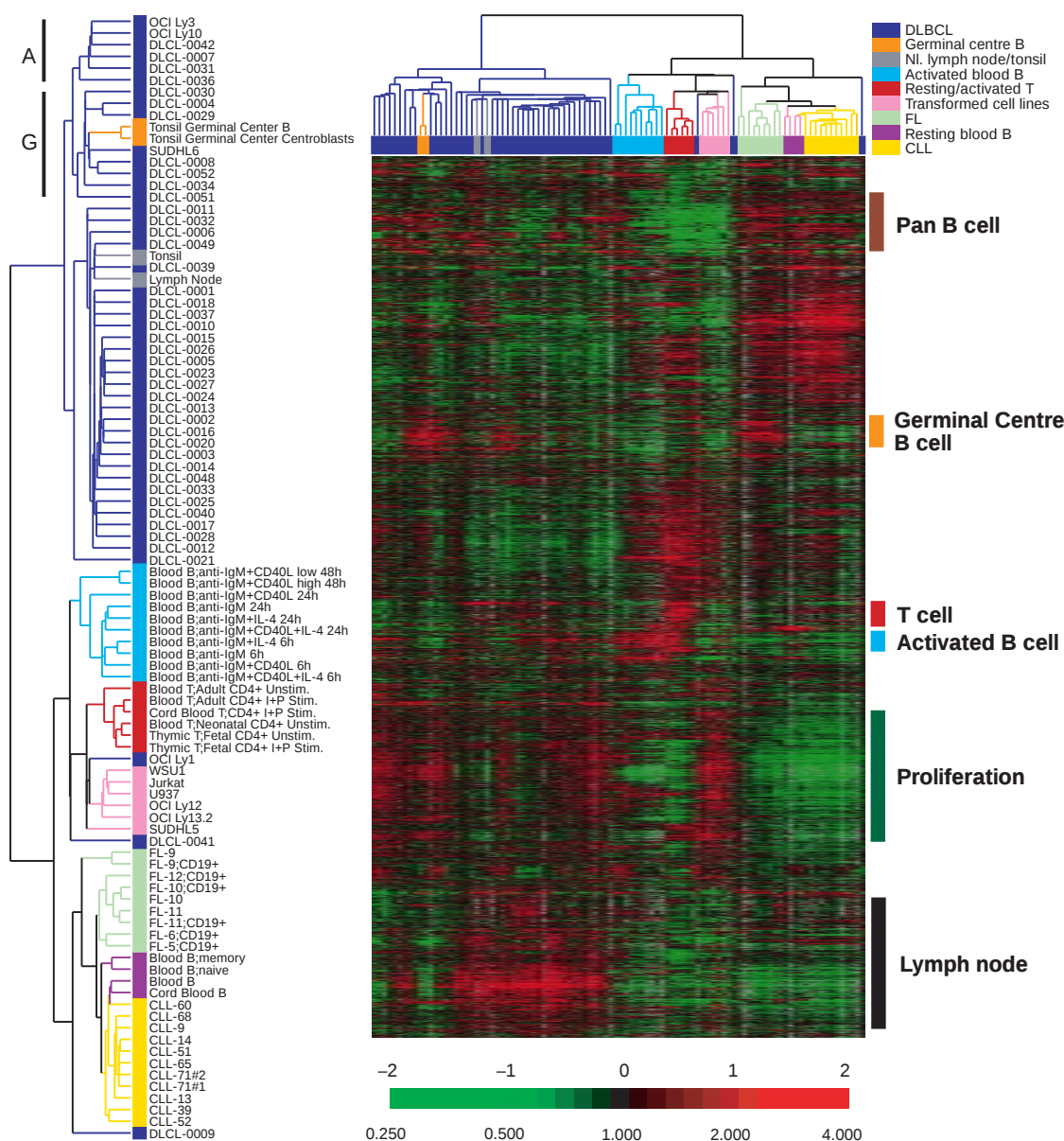


Figure 1 Hierarchical clustering of gene expression data. Depicted are the ~1.8 million measurements of gene expression from 128 microarray analyses of 96 samples of normal and malignant lymphocytes. The dendrogram at the left lists the samples studied and provides a measure of the relatedness of gene expression in each sample. The dendrogram is colour coded according to the category of mRNA sample studied (see upper right key). Each row represents a separate cDNA clone on the microarray and each column a separate mRNA sample. The results presented represent the ratio of

hybridization of fluorescent cDNA probes prepared from each experimental mRNA samples to a reference mRNA sample. These ratios are a measure of relative gene expression in each experimental sample and were depicted according to the colour scale shown at the bottom. As indicated, the scale extends from fluorescence ratios of 0.25 to 4 (–2 to +2 in log base 2 units). Grey indicates missing or excluded data. See Supplementary Information for full data.

The gene expression profiles of DLBCLs were largely distinct from those of CLL and FL and showed additional biological complexity in these biopsy samples. Prominent features of the DLBCL profiles appeared to reflect the non-malignant cells in these tumours. A large group of genes defined a 'lymph-node' signature which was shared by most of the DLBCLs and samples of normal lymph node and tonsil (Fig. 2). This signature featured genes encoding known markers of monocytes and macrophages (CD14, CD105, CSF-1 receptor) and natural killer cells (NK4). In addition, genes involved in the remodelling of the extracellular matrix were abundantly expressed (MMP9 matrix metalloproteinase and TIMP-3). All but one DLBCL biopsy displayed the lymph-node signature, but the intensity of this signature varied, possibly reflecting the relative proportion of tumour and host cells in the lymph-node biopsy.

The variable presence of T lymphocytes in DLBCL biopsies was readily discernible by a T-cell gene expression signature that featured components of the T-cell receptor (TCR- β , CD3 ϵ) and genes downstream of T-cell receptor signalling (fyn, LAT, PKC- θ) (Fig. 2). Although this T-cell expression signature was readily apparent in some DLBCLs, it was virtually undetectable in others.

Discovery of DLBCL subtypes

The structure of the hierarchical dendrogram in Fig. 1 indicated that gene expression patterns in DLBCLs might be inhomogeneous. Three branches of the dendrogram captured most of the DLBCLs with only three outlying samples. Clearly, the position of any given DLBCL sample in the dendrogram is determined in a complicated fashion by the influences of several distinct biological themes that are reflected in the expression pattern. Inspection of the gene expression map shown in Fig. 1 suggested that several independent sets of genes were responsible for much of the DLBCL substructure. The expression signatures related to proliferation, T cells and lymph-node biology were differentially represented in the three DLBCL branches. In addition, we noted that the genes that distinguished germinal centre B cells from other stages in B-cell ontogeny were also differentially expressed among DLBCLs, suggesting that B-cell differentiation genes may also be used to subdivide DLBCL. The expression of the germinal centre B cell genes among DLBCLs varied independently from the expression of genes in the other gene expression signatures (Fig. 2; see Supplementary Information for details). In principle, each of these gene expression

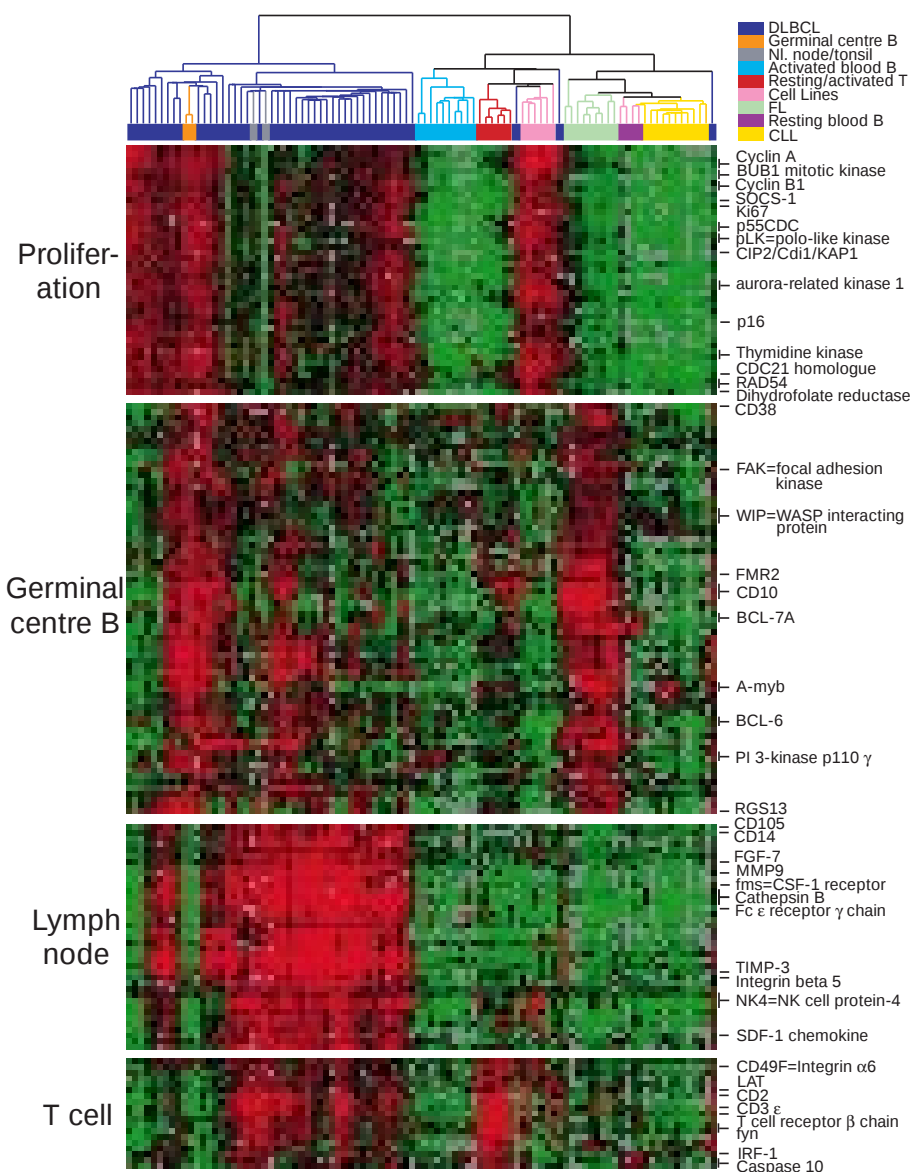


Figure 2 Expanded view of biologically distinct gene expression signatures defined by hierarchical clustering. Data are the same as in Fig. 1. Most genes without designations

on the right are new genes of unknown function derived from various lymphoid cDNA libraries.

signatures could be used to define subsets of DLBCL. We decided to focus our attention initially on the germinal centre B-cell genes, however, because we suspected that these genes might identify DLBCL cases that were derived from distinct stages of normal B-cell differentiation. Indeed, the clustering of the germinal centre B-cell samples with a subset of the DLBCLs in a major branch of the dendrogram in Fig. 1 suggested that this group of DLBCLs might resemble normal germinal centre B cells.

To test this hypothesis, we reclustered the DLBCL cases using only the expression pattern of the genes that define the germinal centre B-cell signature (Fig. 3a). Two large branches were evident in the resulting dendrogram. We will refer to the groups defined by these branches as GC B-like DLBCL and activated B-like DLBCL, for reasons detailed below. The same two branches were also evident in the dendrogram in Fig. 1: activated B-like DLBCL includes all cases in the branch labelled 'A', and GC B-like DLBCL includes all cases in branch labelled 'G'. The largest DLBCL branch in Fig. 1 is a mixture of the cases assigned to the two subgroups. Normal germinal centre B cells were clustered with the GC B-like DLBCL group. Indeed, the DLBCL cases in GC B-like DLBCL group expressed, to a varying degree, all of the genes that define the germinal centre B-cell

signature. In contrast, the activated B-like DLBCL group expressed these genes at low or undetectable levels, for the most part. The gene expression subgroups defined here were not obviously related to histological subtypes of DLBCL: only two of the cases studied could be assigned to the immunoblastic histological subtype, according to the revised Kiel classification system. Furthermore, no evidence of normal germinal centres was found in the lymph-node biopsies. Indeed, one of the germinal centre B-cell markers described below, CD10, was expressed by the lymphoma cells using immunohistochemistry (data not shown). These data clearly suggested that a distinct class of DLBCLs was derived from the germinal centre B cell and retained the gene expression program, and presumably many of the phenotypic characteristics, of this stage of B-cell differentiation.

We searched for genes that were selectively expressed in the activated B-like DLBCL group. This search excluded genes that were readily assigned to the proliferation, T-cell and lymph-node signatures (Fig. 1) in order to focus attention on more subtle intrinsic molecular features of this group of tumours. We used hierarchical clustering to reorder this set of 2,984 genes while maintaining the order shown in Fig. 3a of the DLBCL cases (Fig. 3b). As is evident in Fig. 3c, a cluster of genes could be recognized on the basis of their elevated expression in the activated B-like DLBCLs, as compared with GC B-like DLBCLs. It is important to note that considerable gene expression heterogeneity exists within each subgroup and that no single gene in either of these large clusters was absolutely correlated in expression with the DLBCL subgroup taxonomy. Rather, patients assigned by this method to either DLBCL subgroup shared a large gene expression program that distinguished them from the other subgroup.

DLBCL subgroups and B-cell differentiation

We examined how all of the genes that distinguish these DLBCL subgroups are expressed during B-cell differentiation and activation. Figure 4 shows that almost all of the genes that defined GC B-like DLBCL were highly expressed in normal germinal centre B cells. Most of these genes were expressed at low or undetectable levels in peripheral blood B cells that had been activated *in vitro* by a variety of mitogenic signals. Some of the GC B-like DLBCL genes were expressed in resting blood B cells and germinal centre B cells at comparable levels but not in activated peripheral blood B cells. Conversely, virtually all of the genes that were selectively expressed in germinal centre B cells relative to resting or activated peripheral blood B cells were expressed by GC B-like DLBCL (data not shown).

By contrast, most of the genes that defined activated B-like DLBCL were not expressed in normal germinal centre B cells (Fig. 4). Instead, many of these genes, but not all, were induced during *in vitro* activation of peripheral blood B cells. The time course of expression of these genes during B-cell activation varied, with some genes induced after 6 h of activation and others only expressed after 48 h of activation. Thus, the gene expression signature of activated B-like DLBCLs is reminiscent of, but not identical to, the signature of activated peripheral blood B cells. Notably, two DLBCL cell lines, OCI Ly3 and OCI Ly10, were among the activated B-like DLBCLs. In fact, one or both of these two cell lines expressed virtually all of the genes that defined the activated B-like DLBCL signature. This observation suggests that signal transduction pathways that are inducibly engaged during peripheral B-cell activation and mitogenesis are constitutively active in activated B-like DLBCLs.

The gene expression program that distinguishes GC B-like DLBCLs includes many known markers of germinal centre differentiation (for example, the genes encoding the cell-surface proteins CD10 and CD38 (ref. 24), the nuclear factor A-myb (ref. 25) and the DNA repair protein 8-oxoguanine DNA glycosylase (OGG1)²⁶) and a host of new genes. A particularly noteworthy gene in the GC B-like DLBCL signature is BCL-6, a well-established

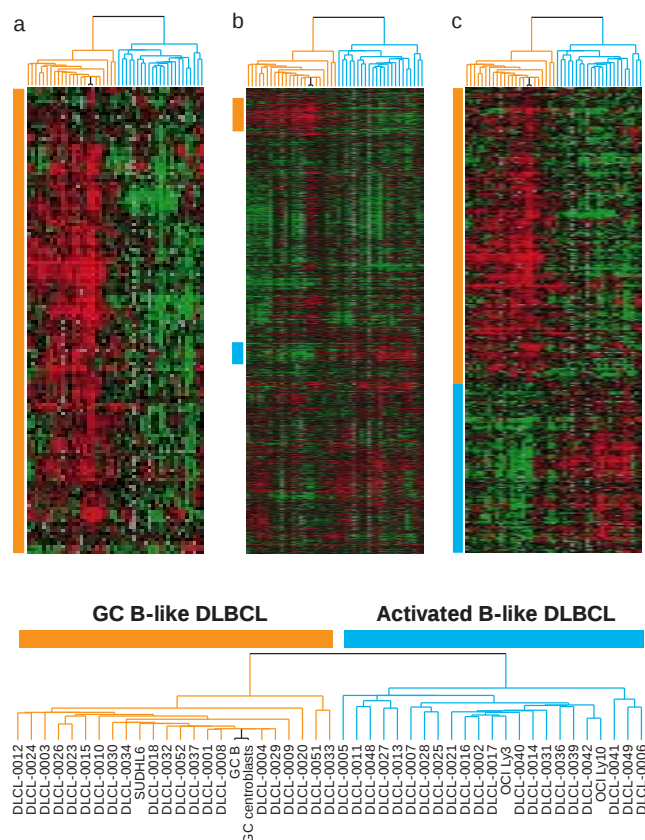


Figure 3 Discovery of DLBCL subtypes by gene expression profiling. The samples used in this clustering analysis are shown at the bottom. **a**, Hierarchical clustering of DLBCL cases (blue and orange) and germinal centre B cells (black) based on the genes of the germinal centre B-cell gene expression signature shown in Figs 1 and 2. Two DLBCL subgroups, GC B-like DLBCL (orange) and activated B-like DLBCL (blue) were defined by this process. **b**, Discovery of genes that are selectively expressed in GC B-like DLBCL and activated B-like DLBCL. All genes from Fig. 1, with the exception of the genes in the proliferation, T-cell and lymph-node gene expression signatures, were ordered by hierarchical clustering while maintaining the order of samples determined in Fig. 3a. Genes selectively expressed in GC B-like DLBCL (orange) and activated B-like DLBCL (blue) are indicated. **c**, Hierarchical clustering of the genes selectively expressed in GC B-like DLBCL and activated B-like DLBCL, which was determined from Fig. 3b.

germinal centre marker that is also the most frequently translocated gene in DLBCL²². Although BCL-6 protein expression is invariably detected in DLBCL, its levels vary and are not correlated with the presence of BCL-6 translocations^{27,28}. Cytogenetic data are available for 16 out of the DLBCL cases studied here and do not support a link between elevated BCL-6 mRNA levels in GC B-like DLBCL and BCL-6 translocations (data not shown). Thus, the higher expression of BCL-6 mRNA in GC B-like DLBCLs is most probably related to their derivation from germinal centre B cells (Fig. 4).

Two other genes that can be altered by translocations in lymphoid malignancies, BCL-7A and LMO2 (TTG-2/RBTN2), have not previously been described as highly expressed in germinal centre

B cells. BCL-7A was cloned as part of a complex chromosomal translocation in a Burkitt's lymphoma cell line and was found to be rearranged in another cell line derived from mediastinal large B-cell lymphoma²⁹. The specific expression of BCL-7A in germinal centre B cells has strong parallels with BCL-6. BCL-6 is required for germinal centre formation during an antigen-driven immune response^{30–32} and is translocated in B-cell malignancies that derive from germinal centre B cells. Given the preferential expression of BCL-7A in germinal centre B cells, it is conceivable that this gene is also involved in normal germinal centre physiology and in the pathophysiology of GC B-like DLBCL. LMO2 is translocated and overexpressed in a subset of T-cell acute lymphoblastic leukaemias

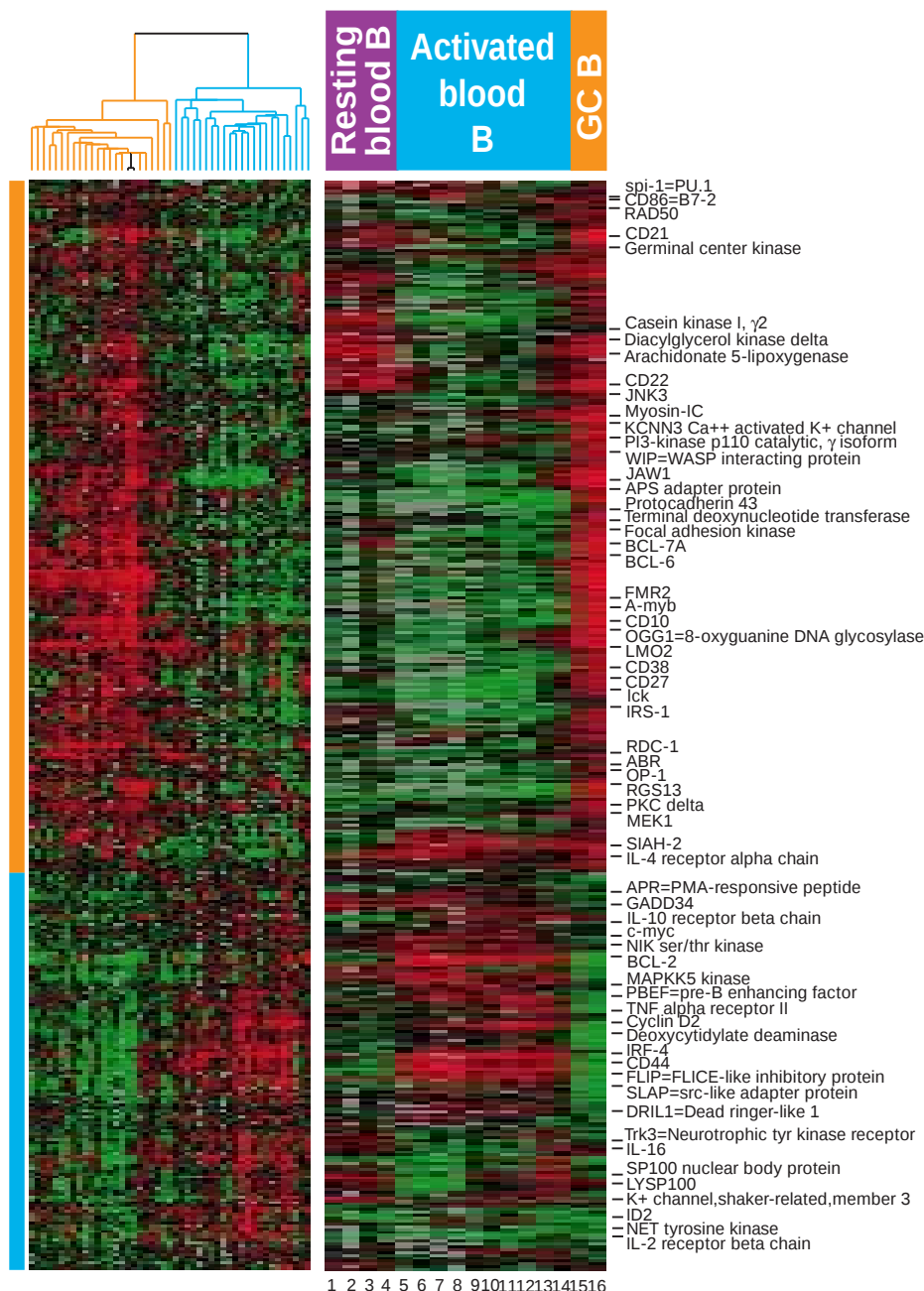


Figure 4 Relationship of DLBCL subgroups to normal B-lymphocyte differentiation and activation. The data in the left panel are taken from Fig. 3c. The right panel depicts gene expression data from the following normal B-cell samples: (1) Total CD19⁺ blood B cells; (2) Naive CD27[−] blood B cells; (3) Memory CD27⁺ blood B cells; (4) cord blood CD19⁺ B cells; (5) blood B cells; anti-IgM 6 h; (6) blood B cells; anti-IgM + IL-4 6 h; (7) blood B cells; anti-IgM + CD40 ligand 6 h; (8) blood B cells; anti-IgM + CD40 ligand + IL-4 6 h; (9) blood

B cells; anti-IgM 24 h; (10) blood B cells; anti-IgM + IL-4 24 h; (11) blood B cells; anti-IgM + CD40 ligand 24 h; (12) blood B cells; anti-IgM + CD40 ligand + IL-4 24 h; (13) blood B cells; anti-IgM + CD40 ligand (low concentration) 48 h; (14) blood B cells; anti-IgM + CD40 ligand (high concentration) 48 h; (15) tonsil germinal centre B cells; (16) tonsil germinal centre centroblasts. See Supplementary Information for full data.

and LMO2 transgenic mice have a block in early T-cell differentiation and develop T-cell leukaemia³³. The selective expression of LMO2 in germinal centre B cells indicates that LMO2 may have a role in inhibiting differentiation in the B-cell lineage as well, and perhaps a corresponding role in the DLBCL malignant phenotype.

The activated B-like DLBCL signature also includes a gene that is translocated in lymphoid malignancies, IRF4 (MUM1/LSIRF). IRF4 is fused to the immunoglobulin locus in some cases of multiple myeloma and can function as an oncogene *in vitro*³⁴. IRF4 is transiently induced during normal lymphocyte activation³⁵ (Fig. 4) and is critical for the proliferation of B lymphocytes in response to signals from the antigen receptor³⁶. Thus, the constitutive expression of IRF4 in activated B-like DLBCLs may contribute to the unchecked proliferation of the malignant cells in these tumours.

A notable feature of the gene expression pattern of activated B-like DLBCLs was the expression of two genes whose products inhibit programmed cell death. FLIP (FLICE-like inhibitory protein/I-FLICE/FLAME-1/Casper/MRIT/CASH/CLARP) is a dominant-negative mimic of caspase 8 (FLICE) which can block apoptosis mediated by Fas and other death receptors³⁷. FLIP is induced early during normal lymphocyte activation, presumably to block activation-induced apoptosis that occurs physiologically later in an immune response. FLIP is highly expressed in many tumour types, and its constitutive expression in activated B-like DLBCLs could inhibit apoptosis of tumour cells induced by host T cells expressing Fas ligand^{38,39}. The key anti-apoptotic gene BCL-2 is translocated in most cases of follicular lymphoma and in a subset of DLBCL. BCL-2 mRNA is not expressed in germinal centre B cells but is induced more than 30-fold during activation of peripheral blood B cells (Fig. 4). Most activated B-like DLBCLs (71%) had BCL-2 mRNA levels more than fourfold higher than were observed in germinal centre B cells (Fig. 4). This overexpression did not correlate with BCL-2 translocations (data not shown). A minority of GC B-like DLBCLs (29%) had similarly elevated BCL-2 mRNA levels, indicating that BCL-2 may also be important in some cases of this DLBCL subgroup.

DLBCL gene expression subgroups define prognostic categories

Does the taxonomy of DLBCL derived from gene expression patterns define clinically distinct subgroups of patients? None of the patients included in this study had been treated before obtaining the biopsy sample. Furthermore, these patients were 'de novo' DLBCL cases that had not obviously arisen from pre-existing low-grade malignancies such as follicular lymphoma. After biopsy, the patients were treated at two medical centres using comparable, standard multi-agent chemotherapy regimens. Figure 5a presents a Kaplan–Meier plot of overall survival data from these patients,

segregated according to gene expression subgroup. Germinal centre B-like and activated B-like DLBCLs were associated with statistically significant differences in overall survival ($P < 0.01$) and in event-free survival (data not shown). Although the average five-year survival for all patients was 52%, 76% of GC B-like DLBCL patients were still alive after five years, as compared with only 16% of activated B-like DLBCL patients. The differential survival of patients in the two DLBCL subgroups was apparently uninfluenced by the anthracycline-based chemotherapeutic regimen used (data not shown), which is not surprising as responses of DLBCL patients to various multi-agent chemotherapeutic regimens were found to be equivalent⁴⁰. Thus, the molecular differences between these two kinds of lymphoma were accompanied by a remarkable divergence in clinical behaviour, suggesting that GC B-like DLBCL and activated B cell DLBCL should be regarded as distinct diseases.

A clinical indicator of prognosis, the International Prognostic Indicator (IPI), has been successfully used to define prognostic subgroups in DLBCL⁸. This indicator takes into account the patient's age, performance status, and the extent and location of disease. As suspected, within our patient population a low IPI score (0–2) identified patients with better overall survival as compared with patients with a high IPI score (3–5) (Fig. 5b). We then determined whether our molecular definition of DLBCL subgroups could add to the prognostic value of this clinical indicator of prognosis. Considering only patients with low clinical risk, as judged by the IPI, patients in the activated B-like DLBCL group had a distinctly worse overall survival than patients in the GC B-like DLBCL group ($P < 0.05$) (Fig. 5c). Thus, the molecular dissection of DLBCL by gene expression profiling and the IPI apparently identify different features of these patients that influence their survival.

Conclusions

This study shows that a genomic view of gene expression in cancer can bring clarity to previously muddy diagnostic categories. The precision of morphological diagnosis, even when supplemented with immunohistochemistry for a few markers, was insufficient in the case of DLBCL to identify believable diagnostic subgroups. A number of individual markers have been used to define subsets of DLBCL^{41–46}, but these studies do not provide the present overview that strongly implies that this single diagnostic category of lymphoma harbours at least two distinct diseases. Indeed, the new methods of gene expression profiling call for a revised definition of what is deemed a 'disease'. The two DLBCL subgroups are distinguished from each other by the differential expression of hundreds of different genes, and these genes relate each subgroup to a separate stage of B-cell differentiation and activation. These molecular differences, in the light of accompanying clinical

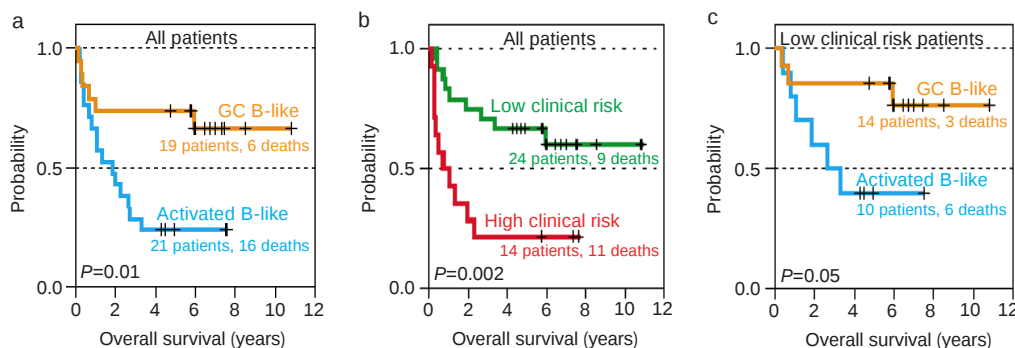


Figure 5 Clinically distinct DLBCL subgroups defined by gene expression profiling. **a**, Kaplan–Meier plot of overall survival of DLBCL patients grouped on the basis of gene expression profiling. **b**, Kaplan–Meier plot of overall survival of DLBCL patients grouped according to the International Prognostic Index (IPI). Low clinical risk patients (IPI score

0–2) and high clinical risk patients (IPI score 3–5) are plotted separately. **c**, Kaplan–Meier plot of overall survival of low clinical risk DLBCL patients (IPI score 0–2) grouped on the basis of their gene expression profiles.

differences between these subgroups, suggest that these two subgroups of DLBCL should be considered separate diseases.

Nonetheless, we do not wish to imply that patients within a DLBCL subgroup defined here are monomorphic. As mentioned above, considerable molecular heterogeneity exists within each DLBCL subgroup. As many more DLBCL patients are studied by gene expression profiling, it is quite possible that more subgroups will emerge. Given that many current diagnostic categories of non-Hodgkin's lymphoma constitute less than 10% of the total cases⁵, it seems likely that the DLBCL diagnostic category will also include a number of minor subgroups.

The classification scheme highlighted in this study divided DLBCL on the basis of genes that are differentially expressed within the B-cell lineage. This particular classification identified patient groups that differed in survival after treatment with anthracycline-based multi-agent chemotherapy regimens. It is unclear at present which of the genes that distinguish GC B-like DLBCL from activated B-like DLBCL are the most important molecular determinants of chemotherapy responsiveness. Furthermore, there is residual clinical heterogeneity which cannot be explained by the current classification. Despite the fact that patients with GC B-like DLBCL had an overall favourable prognosis, five patients died within the first two years of diagnosis. Likewise, three patients in the activated B-like DLBCL subgroup were alive five years after treatment, despite the poor outcome of most patients in this subgroup. By profiling the gene expression of many more DLBCLs, it may become possible to implicate a single gene or pathway in chemotherapy responsiveness with statistical certainty. More probably, however, a multivariate approach to prognosis will be needed that combines knowledge of the DLBCL subgroup, as defined here, with measurements of individual genes or pathways that contribute to treatment outcome.

Gene expression profiling presents a new way of approaching cancer therapeutics in the future. Current treatment of DLBCL typically begins with multi-agent chemotherapy, and then, if a complete remission cannot be maintained, patients are considered for bone marrow transplantation⁷. The definition of prognostic groups by gene expression profiling, in combination with clinical indicators such as the IPI, may lead to the recommendation that some patients receive early bone marrow transplantations upon initial diagnosis. In testing cancer therapeutics in clinical trials, it is obviously beneficial to define homogeneous populations of patients to improve the likelihood of observing efficacy in specific disease entities. We anticipate that global surveys of gene expression in cancer, such as we present here, will identify a small number of marker genes that will be used to stratify patients into molecularly relevant categories which will improve the precision and power of clinical trials.

Finally, the genomic-scale view of gene expression in cancer provides a unique perspective on the development of new cancer therapeutics that could be based on a molecular understanding of the cancer phenotype. Our study shows that the two DLBCL subgroups differentially expressed entire transcriptional modules composed of hundreds of genes, many of which could be expected to contribute to the malignant behaviour of the tumour. This observation suggests that successful new therapeutics might be aimed at the upstream signal-transducing molecules whose constitutive activity in these lymphomas leads to expression of pathological transcriptional programs. □

Methods

Messenger RNA samples

Total germinal centre B cells and centroblasts were purified from human tonsils as described²⁴. Human blood B cells were purified from adult apheresis products or cord blood by magnetic enrichment for CD19⁺ cells (Miltenyi Biotec). Naive CD27⁺ B cells and memory CD27⁺ blood B cells were isolated by fluorescent cell sorting starting with CD19⁺ adult peripheral blood B cells^{47,48}. Magnetic cell sorting was used to purify CD4⁺,

CD45RA^{high} T cells from human cord blood or adult peripheral blood and CD4⁺ thymocytes from human fetal thymus (Miltenyi Biotec). All lymphocyte samples were purified to more than 98% homogeneity as determined by FACS analysis. For rare lymphoid subpopulations such as centroblasts or resting and naive peripheral blood B cells, purified samples from multiple donors were pooled for microarray analysis. *In vitro* stimulation of peripheral B cells was done as described⁴⁹ using anti-IgM antibody, IL-4 and/or CD40 ligand-containing membranes. Most experiments used a 1:1000 dilution of CD40 ligand membranes (designated 'low' concentration, Figs 1 and 4) but one experiment used a 1:200 dilution (designated 'high' concentration, Figs 1 and 4). T cells were stimulated for 2 h with phorbol ester (50 ng ml⁻¹) and ionomycin (1.5 μM). Patient samples were obtained after informed consent and were treated anonymously during microarray analysis. DLBCL patients were treated at either University of Nebraska Medical Center (*n* = 34) or Stanford University School of Medicine (*n* = 8) using comparable, anthracycline-based, multi-agent chemotherapeutic regimens with curative intent. Clinical data were not available on two of the DLBCL cases presented in Fig. 1 (DLCL-51 and DLCL-52). For two additional patients (DLCL-25 and DLCL-36), the data needed to calculate the IPI were not available. DLBCL and FL lymph-node biopsies were either snap frozen, frozen in OCT or disaggregated and frozen as a viable cell suspension. Chronic lymphocyte leukaemia cells were purified from untreated patients by magnetic selection for CD19⁺ cells (Miltenyi Biotec).

Microarray procedures

DNA microarray analysis of gene expression was done essentially as described⁵⁰. The cDNA clones on the Lymphochip microarray are listed in Supplementary Information and are available from Research Genetics. Fluorescent images of hybridized microarrays were obtained using a GenePix 4000 microarray scanner (Axon Instruments). Images were analysed with ScanAlyze (M. Eisen; <http://www.microarrays.org/software>), and fluorescence ratios (along with numerous quality control parameters; see ScanAlyze manual) were stored in a custom database. Single spots or areas of the array with obvious blemishes were flagged and excluded from subsequent analyses. Raw data files for each array containing all measured values and manual flags are available in Supplementary Information. A set of clones that consistently behaved poorly across arrays was identified and excluded from all analyses (see Supplementary Information). Fluorescence ratios were calibrated independently for each array by applying a single scaling factor to all fluorescence ratios from each array; this scaling factor was computed so that the median fluorescence ratio of well-measured spots on each array was 1.0.

All cDNA microarray analyses were performed using poly-(A)⁺ mRNA (Fast Track, Invitrogen). In each experiment, fluorescent cDNA probes were prepared from an experimental mRNA sample (Cy5-labelled) and a control mRNA sample (Cy3-labelled) isolated from a pool of nine lymphoma cell lines (Raji, Jurkat, L428, OCI-Ly3, OCI-Ly8, OCI-Ly1, SUDHL5, SUDHL6 and WSU1). The use of a common control cDNA probe allows the relative expression of each gene to be compared across all samples¹⁸.

Data analysis

All non-flagged array elements for which the fluorescent intensity in each channel was greater than 1.4 times the local background were considered well measured. The ratio values were log-transformed (base 2) and stored in a table (rows, individual cDNA clones; columns, single mRNA samples). Where samples had been analysed on multiple arrays, multiple observations for an array element for a single sample were averaged. Array elements that were not well measured on at least 80% of the 96 mRNA samples were excluded. Data for the remaining genes were centred by subtracting (in log space) the median observed value, to remove any effect of the amount of RNA in the reference pool. This dataset contains 4,026 array elements (see Supplementary Information). Hierarchical clustering was applied to both axes using the weighted pair-group method with centroid average as implemented in the program Cluster (M. Eisen; <http://www.microarrays.org/software>)²⁰. The distance matrices used were Pearson correlation for clustering the arrays and the inner product of vectors normalized to magnitude 1 for the genes (this is a slight variant of Pearson correlation; see Cluster manual available at <http://www.microarrays.org/software/> for computational details). The results were analysed with Tree View (M. Eisen; <http://www.microarrays.org/software>)²⁰. All datasets and image files used to generate Figs 1–4 are included in the Supplementary Information, along with numerous supplementary and additional analyses.

Received 11 November 1999; accepted 10 January 2000.

- Hodgkin, T. On some morbid appearances of the absorbant glands and spleen. *Med.-Chir. Trans.* **17**, 68–114 (1832).
- Sternberg, C. Über eine eigenartige unter dem Bilde der Pseudoleukämie verlaufende Tuberculose des lymphatischen Apparates. *Heilk.* **19**, 21–90 (1898).
- Reed, D. M. On the pathological changes in Hodgkin's disease, with especial reference to its relation to tuberculosis. *Johns Hopkins Hosp. Rep.* **10**, 133–196 (1902).
- Rosenberg, S. A. Classification of lymphoid neoplasms. *Blood* **84**, 1359–1360 (1994).
- Harris, N. L. *et al.* A revised European–American classification of lymphoid neoplasms: a proposal from the International Lymphoma Study Group. *Blood* **84**, 1361–1392 (1994).
- The Non-Hodgkin's Lymphoma Classification Project: A clinical evaluation of the International Lymphoma Study Group classification of non-Hodgkin's lymphoma. *Blood* **89**, 3909–3918 (1997).
- Vose, J. M. Current approaches to the management of non-Hodgkin's lymphoma. *Semin. Oncol.* **25**, 483–491 (1998).
- The International Non-Hodgkin's Lymphoma Prognostic Factors Project: A predictive model for aggressive non-Hodgkin's lymphoma. *N. Engl. J. Med.* **329**, 987–994 (1993).
- Klein, U. *et al.* Somatic hypermutation in normal and transformed human B cells. *Immunol. Rev.* **162**, 261–280 (1998).

10. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
11. Bubendorf, L. *et al.* Hormone therapy failure in human prostate cancer: analysis by complementary DNA and tissue microarrays. *J. Natl Cancer Inst.* **91**, 1758–1764 (1999).
12. Wang, K. *et al.* Monitoring gene expression profile changes in ovarian carcinomas using cDNA microarray. *Gene* **229**, 101–108 (1999).
13. Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
14. Khan, J. *et al.* Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res.* **58**, 5009–5013 (1998).
15. Perou, C. M. *et al.* Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl Acad. Sci. USA* **96**, 9212–217 (1999).
16. DeRisi, J. *et al.* Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genet.* **14**, 457–460 (1996).
17. Alon, U. *et al.* Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA* **96**, 6745–6750 (1999).
18. Alizadeh, A. *et al.* The Lymphochip: a specialized cDNA microarray for the genomic-scale analysis of gene expression in normal and malignant lymphocytes. *Cold Spring Harbor Symp. Quant. Biol.* (in the press).
19. Alizadeh, A., Eisen, M., Botstein, D., Brown, P. O. & Staudt, L. M. Probing lymphocyte biology by genomic-scale gene expression analysis. *J. Clin. Immunol.* **18**, 373–379 (1998).
20. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
21. Grogan, T. M. *et al.* Independent prognostic significance of a nuclear proliferation antigen in diffuse large cell lymphomas as determined by the monoclonal antibody Ki-67. *Blood* **71**, 1157–1160 (1988).
22. Staudt, L. M., Dent, A. L., Shaffer, A. L. & Yu, X. Regulation of lymphocyte cell fate decisions and lymphomagenesis by BCL-6. *Int. J. Immunol.* **18**, 381–403 (1999).
23. Bahler, D. W. & Levy, R. Clonal evolution of a follicular lymphoma: evidence for antigen selection. *Proc. Natl Acad. Sci. USA* **89**, 6770–6774 (1992).
24. Liu, Y. -J. & Banchereau, J. in *Handbook of Experimental Immunology* (eds Weir, D., Blackwell, C., Herzenberg, L. & Herzenberg, L.) 93.1–93.9 (Blackwell Scientific, Oxford, 1996).
25. Golay, J., Erba, E., Bernasconi, S., Peri, G. & Introna, M. The A-myb gene is preferentially expressed in tonsillar CD38+, CD39–, and sIgM– B lymphocytes and in Burkitt's lymphoma cell lines. *J. Immunol.* **153**, 543–553 (1994).
26. Kuo, F. C. & Sklar, J. Augmented expression of a human gene for 8-oxoguanine DNA glycosylase (MutM) in B lymphocytes of the dark zone in lymph node germinal centers. *J. Exp. Med.* **186**, 1547–1556 (1997).
27. Flenghi, L. *et al.* A specific monoclonal antibody (PG-B6) detects expression of the BCL-6 protein in germinal center B cells. *Am. J. Pathol.* **147**, 405–411 (1995).
28. Pittaluga, S. *et al.* BCL-6 expression in reactive lymphoid tissue and in B-cell non-Hodgkin's lymphomas. *J. Pathol.* **179**, 145–150 (1996).
29. Zani, V. J. *et al.* Molecular cloning of complex chromosomal translocation t(8;14;12)(q24.1;q32.3;q24.1) in a Burkitt lymphoma cell line defines a new gene (BCL7A) with homology to caldesmon. *Blood* **87**, 3124–3134 (1996).
30. Fukuda, T. *et al.* Disruption of the Bcl6 gene results in an impaired germinal center formation. *J. Exp. Med.* **186**, 439–448 (1997).
31. Ye, B. H. *et al.* The BCL-6 proto-oncogene controls germinal-centre formation and Th2-type inflammation. *Nature Genet.* **16**, 161–170 (1997).
32. Dent, A. L., Shaffer, A. L., Yu, X., Allman, D. & Staudt, L. M. Control of inflammation, cytokine expression, and germinal center formation by BCL-6. *Science* **276**, 589–592 (1997).
33. Rabbitts, T. H. LMO T-cell translocation oncogenes typify genes activated by chromosomal translocations that alter transcription and developmental processes. *Genes Dev.* **12**, 2651–2657 (1998).
34. Iida, S. *et al.* Deregulation of MUM1/IRF4 by chromosomal translocation in multiple myeloma. *Nature Genet.* **17**, 226–230 (1997).
35. Matsuyama, T. *et al.* Molecular cloning of LSIRF, a lymphoid-specific member of the interferon regulatory factor family that binds the interferon-stimulated response element (ISRE). *Nucleic Acids Res.* **23**, 2127–2136 (1995).
36. Mittrucker, H. W. *et al.* Requirement for the transcription factor LSIRF/IRF4 for mature B and T lymphocyte function. *Science* **275**, 540–543 (1997).
37. Tschopp, J., Irmeler, M. & Thome, M. Inhibition of fas death signals by FLIPs. *Curr. Opin. Immunol.* **10**, 552–558 (1998).
38. Djerbi, M. *et al.* The inhibitor of death receptor signaling, FLICE-inhibitory protein defines a new class of tumor progression factors. *J. Exp. Med.* **190**, 1025–1031 (1999).
39. Medema, J. P., de Jong, J., van Hall, T., Melief, C. J. M. & Offringa, R. Immune escape of tumors *in vivo* by expression of cellular FLICE-inhibitory protein. *J. Exp. Med.* **190**, 1033–1038 (1999).
40. Fisher, R. I. *et al.* Comparison of a standard regimen (CHOP) with three intensive chemotherapy regimens for advanced non-Hodgkin's lymphoma. *N. Engl. J. Med.* **328**, 1002–1006 (1993).
41. Jalkanen, S., Joensuu, H., Soderstrom, K. O. & Kleml, P. Lymphocyte homing and clinical behavior of non-Hodgkin's lymphoma. *J. Clin. Invest.* **87**, 1835–1840 (1991).
42. Harada, S. *et al.* Molecular and immunological dissection of diffuse large B cell lymphoma: CD5–, and CD5+ with CD10+ groups may constitute clinically relevant subtypes. *Leukemia* **13**, 1441–1447 (1999).
43. Kramer, M. H. *et al.* Clinical significance of bcl2 and p53 protein expression in diffuse large B-cell lymphoma: a population-based study. *J. Clin. Oncol.* **14**, 2131–2138 (1996).
44. Preti, H. A. *et al.* Prognostic value of serum interleukin-6 in diffuse large-cell lymphoma. *Ann. Int. Med.* **127**, 186–194 (1997).
45. Gascoyne, R. D. *et al.* Prognostic significance of Bcl-2 protein expression and Bcl-2 gene rearrangement in diffuse aggressive non-Hodgkin's lymphoma. *Blood* **90**, 244–251 (1997).
46. Kramer, M. H. *et al.* Clinical relevance of BCL2, BCL6, and MYC rearrangements in diffuse large B-cell lymphoma. *Blood* **92**, 3152–3162 (1998).
47. Klein, U., Rajewsky, K. & Kuppers, R. Human immunoglobulin (Ig)M+IgD+ peripheral blood B cells expressing the CD27 cell surface antigen carry somatically mutated variable region genes: CD27 as a general marker for somatically mutated (memory) B cells. *J. Exp. Med.* **188**, 1679–1689 (1998).
48. Tangye, S. G., Liu, Y. J., Aversa, G., Phillips, J. H. & de Vries, J. E. Identification of functional human splenic memory B cells by expression of CD148 and CD27. *J. Exp. Med.* **188**, 1691–1703 (1998).
49. Allman, D. *et al.* BCL-6 expression during B-cell activation. *Blood* **87**, 5257–5268 (1996).
50. Eisen, M. B. & Brown, P. O. DNA arrays for analysis of gene expression. *Methods Enzymol.* **303**, 179–205 (1999).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>), on the authors' World-Wide Web site (<http://llmpp.nih.gov/lymphoma>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We acknowledge the support of the Cancer Genome Anatomy Project (CGAP), led by B. Strausberg and R. Klausner. We also thank R. Klausner for comments on the manuscript; C. Prange for providing CGAP cDNA clones; H. Messner for providing DLBCL cell lines; H. Mostowski for sorting lymphocyte subpopulations by FACS; Holy Cross Hospital, Silver Spring, Maryland, for providing human tonsils; J. DeRisi for helpful advice on microarray technology; and members of the Staudt, Brown and Botstein laboratories for helpful discussions. Research at Stanford was supported by grants from the National Cancer Institute to D.B., R.L. and P.O.B. and by the Howard Hughes Medical Institute. P.O.B. is an Associate Investigator of the Howard Hughes Medical Institute. A.A. was initially supported by the Howard Hughes Medical Institute Research Scholar Program while at the NIH and then by the Medical Scientist Training Program at Stanford University. M.B.E. was supported by a Computational Molecular Biology Postdoctoral Fellowship from the Alfred E. Sloan Foundation.

Correspondence and requests for materials should be addressed to L.M.S. (lstaedt@box-1.nih.gov) or P.O.B. (pbrown@cmgm.stanford.edu).

Quantum mirages formed by coherent projection of electronic structure

H. C. Manoharan, C. P. Lutz & D. M. Eigler

IBM Research Division, Almaden Research Center, 650 Harry Road, San Jose, California 95120, USA

Image projection relies on classical wave mechanics and the use of natural or engineered structures such as lenses or resonant cavities. Well-known examples include the bending of light to create mirages in the atmosphere, and the focusing of sound by whispering galleries. However, the observation of analogous phenomena in condensed matter systems is a more recent development¹, facilitated by advances in nanofabrication. Here we report the projection of the electronic structure surrounding a magnetic Co atom to a remote location on the surface of a Cu crystal; electron partial waves scattered from the real Co atom are coherently refocused to form a spectral image or 'quantum mirage'. The focusing device is an elliptical quantum corral^{2,3}, assembled on the Cu surface. The corral acts as a quantum mechanical resonator, while the two-dimensional Cu surface-state electrons form the projection medium. When placed on the surface, Co atoms display a distinctive spectroscopic signature, known as the many-particle Kondo resonance^{4–6}, which arises from their magnetic moment. By positioning a Co atom at one focus of the ellipse, we detect a strong Kondo signature not only at the atom, but also at the empty focus. This behaviour contrasts with the usual spatially-decreasing response of an electron gas to a localized perturbation⁷.

Experiments were performed with a scanning tunnelling microscope (STM) operating at 4 K in ultrahigh vacuum (UHV). The (111) surface of a single-crystal Cu sample was prepared by several cycles of Ar ion sputtering and UHV annealing, with surface cleanliness monitored by Auger spectroscopy. The sample was then cooled to 4 K and dosed with a calibrated Co electron-beam evaporator to typical coverages of ~ 13 atoms $(100 \text{ \AA})^{-2}$ (0.007 monolayer). We use the convention that the voltage V across the STM tunnel junction is the sample bias with respect to the tip. As the STM tip, a polycrystalline Ir wire, was prepared through field emission followed by gentle collisions with the Cu sample, the exact chemical identity of the last atom terminating the tip was unknown. However, we have repeated the spectroscopy results given here with a tip terminated in an atom of known identity (Xe) to confirm that our results are independent of tip chemistry.

We demonstrate the existence of quantum mirages by using the Kondo resonance⁴, a response which forms around individual magnetic moments on the metal surface. Although it was identified more than 30 years ago and studied extensively since then⁸, the prototypical Kondo problem—namely, a single, well-characterized, magnetic impurity embedded in a non-magnetic host conductor—has eluded experimental access until quite recently, when STM spectroscopy was used to identify the Kondo resonance around single magnetic atoms placed on a non-magnetic metal surface^{5,6}. The Kondo effect arises as a result of the exchange interaction between the localized magnetic moment of the impurity and the delocalized electrons in the metallic host. At sufficiently low temperatures T , nearby conduction electrons begin to align their spins to screen the spin on the local moment. A many-body spin-singlet state is thus formed, consisting of the local magnetic impurity surrounded by a spin compensation cloud of conduction electrons, such that no net moment remains at the Kondo site. This intriguing state has a special energy scale $k_B T_K$ associated with it, where T_K is the Kondo temperature (k_B is the Boltzmann constant). Only for $T < T_K$, and for certain combinations of magnetic atoms

and host metals, is the Kondo effect expected to exist. As the constituent electrons in the Kondo screening cloud come predominantly from the Fermi surface of the metallic host, the Kondo effect may be observed spectroscopically as an acute perturbation or resonance in the renormalized density of states near the Fermi energy E_F . The spectral width of this resonance scales proportionally with T_K .

In Fig. 1 we present tunnelling spectra of Co atoms on Cu(111), in which the Kondo resonance is manifest as a sharp suppression in differential conductance dI/dV in the immediate vicinity of E_F ($V = 0$). As shown in Fig. 1a, this resonance is spatially centred precisely on the Co atoms and dies off over a lateral length scale of $\sim 10 \text{ \AA}$. The observed spectroscopic feature is quite narrow (9 mV full-width at half-maximum, FWHM) and very strong (the resonance magnitude is about 40% of the zero-bias conductance). The shift in the resonance minimum from $V = 0$ is not a measurement artifact and is determined to be about 1 mV. We observe nearly identical resonance features on all Co atoms encountered and with a variety of different STM tips.

Previously observed resonances associated with the Kondo effect over single magnetic atoms^{5,6} have been interpreted in terms of Fano's model of interfering discrete and continuum channels⁹. We

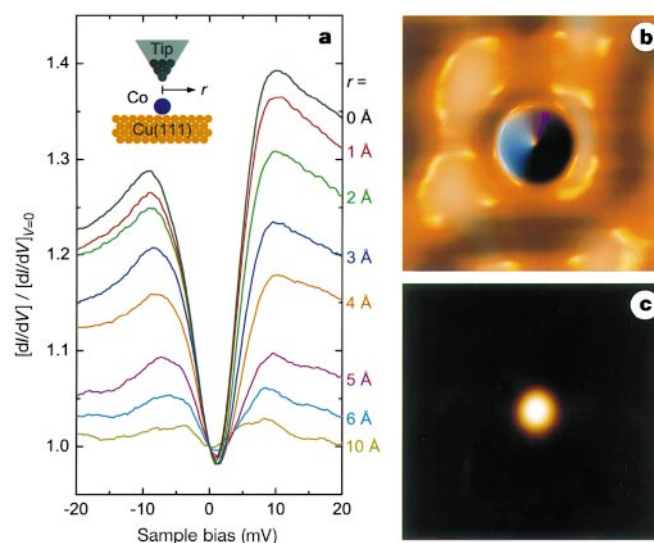


Figure 1 Detection of the Kondo resonance localized around a single Co atom on Cu(111). **a**, Tunnel spectra (normalized dI/dV) acquired over the Co atom for increasing lateral displacement r ($R_T = 100 \text{ M}\Omega$ at $V = 10 \text{ mV}$). Inset, measurement geometry. **b**, 35-Å square topograph ($V = 5 \text{ mV}$, $I = 1 \text{ nA}$) of an isolated Co atom (0.8-Å-high central bump). **c**, dI/dV map of the same area (average of $V = \pm 5 \text{ mV}$ acquisitions, $V_{a.c.} = 1 \text{ mV r.m.s.}$, $I = 1 \text{ nA}$). Dark to light corresponds to increasing conductance. Examples of the three types of data obtained in this experiment: (1) Topograph images (**b**) were acquired with the scanning tunnelling microscope (STM) operating in constant d.c. current (I) mode, in which a closed feedback loop constantly adjusted tip height. (2) Tunnel spectra (**a**) were acquired by adding a small a.c. modulation $V_{a.c.}$ (1 mV r.m.s. at 201 Hz) to the d.c. bias V , opening the feedback loop (hence, holding the tip motionless with respect to the surface), and measuring dI/dV versus V through lock-in detection of the a.c. component of the tunnel current. Such spectra were essentially independent of the tunnel junction impedance R_T , determined by V/I before opening the feedback loop. (3) dI/dV image maps (**c**) were acquired simultaneously with associated topographs by applying an a.c. modulation (typically 250 μV to 1 mV r.m.s. at 201 or 1007 Hz) at a frequency exceeding the bandwidth of the feedback loop, and recording the lock-in detected dI/dV (conductance map) along with tip height (topograph) at fixed d.c. bias V while the tip was scanned in closed-loop constant- I mode. Both kinds of differential conductance measurements constitute a probe of the local density of states under the tip^{17,18}.

find that our observed spectra can be fitted very well by a Fano line shape in the limit of small coupling to the discrete state ($q \ll 1$ in ref. 9); such fits reproduce the 'dip' structure along with the observed asymmetry and the shift in the minimum from $V = 0$. Following the theoretical results of ref. 6, the half width of the Fano resonance ($\Gamma/2$ in ref. 9) is taken to be equal to $k_B T_K$. Fits to our data from a large sampling of Co atoms on the Cu(111) surface yield $T_K = 53 \pm 5$ K.

The Cu(111) surface harbours a two-dimensional free electron gas. These electrons inhabit a surface state band which starts 450 mV below E_F . Co atoms placed on the surface are immersed in the two-dimensional electron sea. This fact is crucial to our results, as it is precisely these electrons that form the projection medium for the quantum mirage. In the topograph of Fig. 1b, the standing waves (equivalently, the energy-resolved Friedel oscillations) formed from the two-dimensional electrons scattering off a single Co atom are evident^{10,11}. Weak variations in the Kondo resonance as a particular Co atom is moved along the surface correlate with the features in the underlying standing-wave field.

We are able to spatially map the Kondo screening cloud by simultaneously acquiring a dI/dV image while V is tuned to a small bias V_0 near the edge of the resonance ($|V_0|$ ranges from 5 to 10 mV for the measurements described). The presence of the Kondo resonance at a particular location on the surface removes spectral density between $V = 0$ and $V = V_0$ and, because the data is acquired in closed-loop constant-current mode, causes the tip to approach the sample and thus increases the measured dI/dV signal.

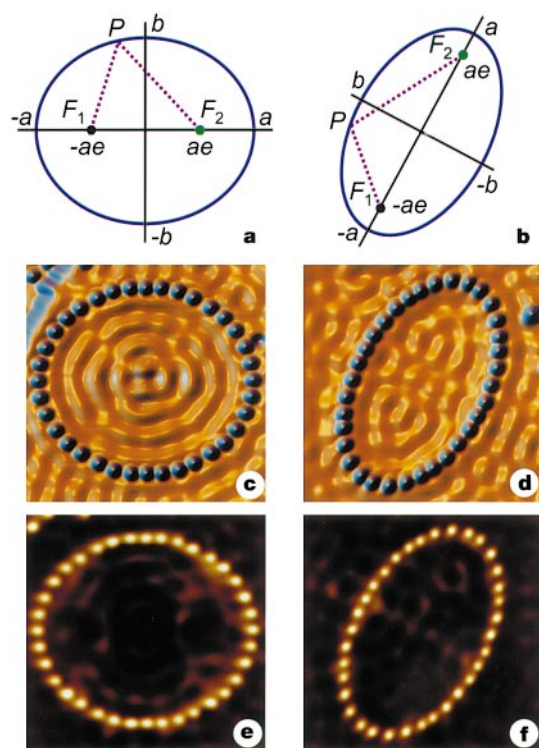


Figure 2 Elliptical electron resonators. **a**, Eccentricity $e = 1/2$; **b**, $e = 0.786$. **c**, **d**, Corresponding topographs of the experimental realizations ($a = 71.3$ Å for both ellipses) employing Co atoms to corral two-dimensional electrons on Cu(111). **e**, **f**, dI/dV maps acquired simultaneously with the corresponding topographs, tuned to image the Kondo resonance. ($V = 10$ mV, $V_{a.c.} = 250$ μ V r.m.s., $I = 1$ nA for **c** and **e**; $V = 8$ mV, $V_{a.c.} = 1$ mV r.m.s., $I = 1$ nA for **d** and **f**). Image dimensions are 150 Å square and 154 Å square for the $e = 1/2$ and $e = 0.786$ ellipses, respectively.

Figure 1c shows the dI/dV map associated with the single Co atom topograph in Fig. 1b. This visualization clearly reveals the location and extent of the Kondo resonance localized at the Co atom site.

We have used the discrete electronic states of elliptical quantum corrals^{2,3} to project the mirage. As illustrated in Fig. 2a and b, an ellipse has the property that all classical paths emanating from one focus F_1 bounce specularly off the wall at arbitrary point P and converge on the second focus F_2 . Furthermore, this path length (dashed line in Fig. 2a and b) remains fixed at $F_1 P F_2 = 2a$ independent of the trajectory, where a is the semimajor axis length. Hence, if a scatterer is placed at one focus, all scattered waves will add with coherent phase at the other focus. We constructed a pair of resonators with the geometries sketched in Fig. 2a

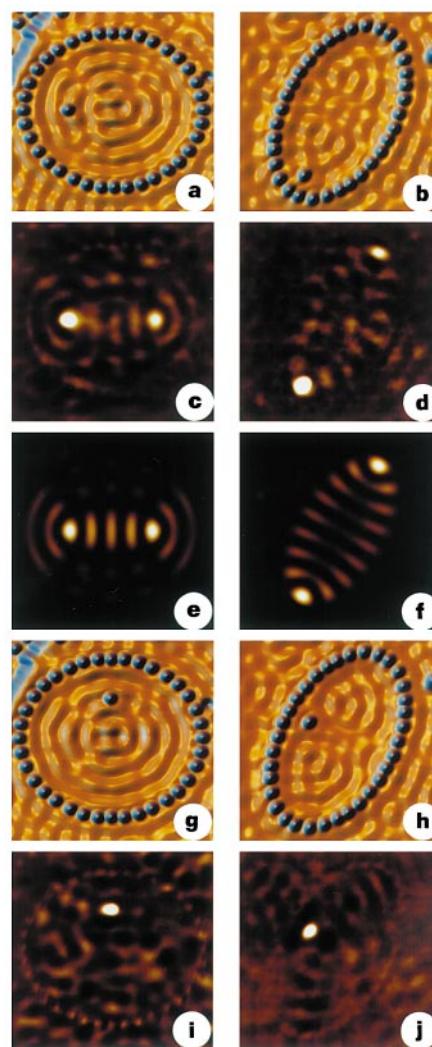


Figure 3 Visualization of the quantum mirage. **a**, **b**, Topographs showing the $e = 1/2$ (**a**) and $e = 0.786$ (**b**) ellipse each with a Co atom at the left focus. **c**, **d**, Associated dI/dV difference maps showing the Kondo effect projected to the empty right focus, resulting in a Co atom mirage. **e**, **f**, Calculated eigenmodes at E_F (magnitude of the wavefunction is plotted). When the interior Co atom is moved off focus (**g** and **h**, topographs), the mirage vanishes (**i** and **j**, corresponding dI/dV difference maps). Imaging conditions and dimensions as in Fig. 2. We have assembled over 20 elliptical resonators of varying size and eccentricity and searched for the formation of a quantum mirage. We find that as a is increased monotonically while e is fixed, the mirage is switched on and off. In each period of this switching, the classical path length $2a$ changes by a half Fermi wavelength. Because we also observe that two focal atoms, one on each focus, couple quite strongly with one another (as judged by the perturbation of the Kondo resonance) these oscillatory results may have a source akin to the RKKY (indirect exchange) interaction¹⁹.

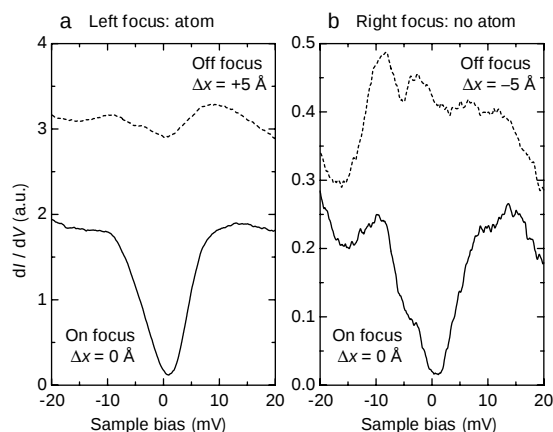


Figure 4 Tunnelling spectra. **a, b**, Near the left (**a**) and near the right (**b**) foci of the $e = 1/2$ elliptical resonator in the configuration depicted in Fig. 3a and c ($R_T = 100 \text{ M}\Omega$ at $V = 10 \text{ mV}$). Solid traces were acquired directly on focus; dashed traces were acquired with the tip displaced $5 \text{ \AA}$ off focus toward the centre of the ellipse. The electronic

structure localized around the Co atom at the left focus is transmitted to the unoccupied right focus. (Note the axis scale change showing $\sim 8 \times$ signal attenuation. A background of constant slope has been subtracted from the data.)

and **b** by using the adatom sliding process^{12,13} to position Co atoms to form the corral walls. STM topographs of the assembled resonators are shown in Fig. 2c and d, where the electron-confining effects are evident. The mean spacing between the Co atoms comprising the resonator walls is roughly four atomic sites on the underlying Cu(111) lattice, which has a nearest-neighbour spacing of $2.55 \text{ \AA}$ at $T = 4 \text{ K}$. This pair of ellipses was chosen to share the same $a = 71.3 \text{ \AA}$ (hence the same focus-to-focus path length) but to have significantly different eccentricity $e = \sqrt{1 - b^2/a^2}$ (where b is the semiminor axis). The left and right columns of Fig. 2 pertain to $e = 1/2$ and $e = 0.786$ ellipses, which corral 84 and 56 electrons, respectively.

The associated dI/dV maps of the empty resonators are displayed in Fig. 2e and f. These maps show the Kondo resonance localized around each Co wall atom, but lack any significant Kondo signal within the ellipse. The faint interior features derive from the local density of states (LDOS) undulations of the ellipse eigenmodes closest to E_F . From these maps it is clear that the wall atoms themselves are not projecting the observed Kondo effect (Fig. 1a) into the confines of the resonator.

Next we studied the effects of placing a single Co atom at various locations inside each resonator. The results are summarized in Fig. 3. For a given geometry with an interior atom, we again acquired a simultaneous topograph and dI/dV map, and then subtracted the respective dI/dV map shown in Fig. 2e and f in order to remove the background electronic contribution of the standing wave LDOS and emphasize the Kondo component of the signal. The resulting dI/dV difference map was therefore tuned to spatially locate the Kondo signature of Co inside the resonator, while the associated topograph revealed the actual atom locations. The striking results of this measurement, shown in Fig. 3a–d, reveal two interior positions at which the Kondo resonance dominates for each ellipse: one signal is localized on the real Co atom at the left focus, while the other signal is centred on the empty right focus. In effect, a phantom (albeit weaker) copy of the Co atom has been created by projecting the localized electronic structure from the occupied focus to the unoccupied focus, thus creating the quantum mirage.

Since we are operating in a regime of reasonably large quantum numbers (~ 30 – 40), the most direct application of semiclassical concepts¹⁴ such as classical foci, path lengths, s -wave scattering, and so on, can help explain some of the effects observed¹⁵. However, also visible in the dI/dV difference maps of Fig. 3c and d is reproducible fine structure throughout the interior of the resonators. Further

insight into these observations comes from treating the system as strictly quantized. Calculations reveal that the additional features in the conductance maps correspond very well to the spatial structure of the eigenstate closest to E_F (approximating the elliptical boundary as a hard-wall box). The magnitude of the corresponding eigenstate for each ellipse is plotted in Fig. 3e and f. The quantum mirage is evidently projected predominantly through this state, which acts as a conduit between foci for the Kondo signature.

Modified geometries show that if the Co atom is placed at other positions within the ellipse besides one of the foci, even at high-symmetry points as shown in Fig. 3g and h, a corresponding mirage is not observed (see Fig. 3i and j).

The Kondo signature detected at the unoccupied focus (see Fig. 3c and d) exhibits about a third of the strength of that at the focal Co atom, while possessing comparable spatial extent. A component of this signal arises from the simple perturbation of the eigenmodes pictured in Fig. 3e and f, as their energies are shifted owing to an impurity inhabiting a region of high electron probability. For this reason, it is important also to examine standard open-loop tunnelling spectra acquired over the Co atom and its associated mirage.

In Fig. 4 we show the results of this test applied to the $e = 1/2$ ellipse. We have measured dI/dV near the left occupied focus (Fig. 4a) to compare with the corresponding spectra near the right unoccupied focus (Fig. 4b). The lower solid traces, obtained directly on focus, demonstrate that the mirage at the right focus is a surprisingly faithful spectroscopic replica of the real atom at the left focus: the resonance line shapes, widths, and zero-bias shifts are equivalent, and the Kondo signal is essentially only attenuated (by a factor of about eight). Moving off focus from either the real or phantom atom by the same distance ($5 \text{ \AA}$) toward the centre of the ellipse (upper dashed curves in Fig. 4) causes the detected Kondo signature to rapidly collapse, in agreement with the dI/dV difference image in Fig. 3c. These results provide perhaps the clearest demonstration that the electronic structure surrounding a focal Co atom is projected intact to the opposite focus.

We have confirmed that non-magnetic scatterers (sulphur and carbon monoxide) placed at one focus neither exhibit nor transmit a Kondo resonance to the opposite focus. We have also substituted non-magnetic carbon monoxide molecules for the Co wall atoms to verify that removing the Kondo effect along the wall does not influence the essential physics of the quantum mirage.

Some possible applications of the observed phenomena are evident. The elliptic geometry coupled with two-dimensional

electrons could be employed to probe atoms and molecules remotely while minimizing the perturbing influences of a nearby local probe—these unwanted interactions might be chemical in nature or result from electric or magnetic fields emanating from an STM tip or other probe device. Our results also suggest altered geometries such as ellipsoids confining bulk electrons, or specially shaped electron mirrors (such as parabolic reflectors) electronically coupling two or more points. One might additionally envision performing other types of ‘spectroscopy-at-a-distance’ beyond electronic structure measurements, for example, detecting vibrational¹⁶ or magnetic excitations.

We conclude with some remaining questions. Our calculations show that ellipses have a class of eigenmodes that possess strongly peaked probability amplitude near the classical foci (Fig. 3e and f are good examples). Our experiments reveal that the strongest mirages occur when one of these eigenmodes is at E_F and therefore at the energy of the Kondo resonance. When perturbed by a focal atom, this eigenmode still has substantial density surrounding both foci; it is therefore plausible that this quantum state ‘samples’ the Kondo resonance on the real atom and transmits that signal to the other focus. The physics behind this process is not completely understood. Also, given that a Kondo signature is detected at the empty focus, does this imply that a measurement there is simply providing us with a remote probe of the Co atom with the intervening two-dimensional electrons acting essentially like a wire? Or, as the Kondo effect on the Co atom requires a net spin polarization of the surrounding electron gas, does the projection of the Kondo resonance imply a modified spin polarization at the empty focus? We speculate that both interpretations are actually correct. Full answers to these questions, however, await further experimental and theoretical efforts.

Received 12 October; accepted 14 December 1999.

1. Spector, J., Stormer, H. L., Baldwin, K. W., Pfeiffer, L. N. & West, K. W. Electron focusing in two-dimensional systems by means of an electrostatic lens. *Appl. Phys. Lett.* **56**, 1290–1292 (1990).
2. Crommie, M. F., Lutz, C. P. & Eigler, D. M. Confinement of electrons to quantum corrals on a metal surface. *Science* **262**, 218–220 (1993).
3. Heremans, J. J., von Molnár, S., Awschalom, D. D. & Gossard, A. C. Ballistic electron focusing by elliptic reflecting barriers. *Appl. Phys. Lett.* **74**, 1281–1283 (1999).
4. Kondo, J. Resistance minimum in dilute magnetic alloys. *Prog. Theor. Phys.* **32**, 37–49 (1964).
5. Li, J., Schneider, W.-D., Berndt, R. & Delley, B. Kondo scattering observed at a single magnetic impurity. *Phys. Rev. Lett.* **80**, 2893–2896 (1998).
6. Madhavan, V., Chen, W., Jamneala, T., Crommie, M. F. & Wingreen, N. S. Tunneling into a single magnetic atom: Spectroscopic evidence of the Kondo resonance. *Science* **280**, 567–569 (1998).
7. Kittel, C. *Quantum Theory of Solids* (Wiley, New York, 1963).
8. Hewson, A. C. *The Kondo Problem to Heavy Fermions* (Cambridge Univ. Press, Cambridge, 1997).
9. Fano, U. Effects of configuration interaction on intensities and phase shifts. *Phys. Rev.* **124**, 1866–1878 (1961).
10. Crommie, M. F., Lutz, C. P. & Eigler, D. M. Imaging standing waves in a two-dimensional electron gas. *Nature* **363**, 524–527 (1993).
11. Hasegawa, Y. & Avouris, P. Direct observation of standing wave formation at surface steps using scanning tunneling spectroscopy. *Phys. Rev. Lett.* **71**, 1071–1074 (1993).
12. Eigler, D. M. & Schweizer, E. K. Positioning single atoms with a scanning tunnelling microscope. *Nature* **344**, 524–526 (1990).
13. Strosio, J. A. & Eigler, D. M. Atomic and molecular manipulation with the scanning tunneling microscope. *Science* **254**, 1319–1326 (1991).
14. Tomsovic, S. & Heller, E. J. Semiclassical construction of chaotic eigenstates. *Phys. Rev. Lett.* **70**, 1405–1408 (1993).
15. Chan, Y. S. & Heller, E. J. Scanning tunnel microscopy surface state electron scattering: Two-tip results from one-tip data. *Phys. Rev. Lett.* **78**, 2570–2572 (1997).
16. Stipe, B. C., Rezaei, M. A. & Ho, W. Single-molecule vibrational spectroscopy and microscopy. *Science* **280**, 1732–1735 (1998).
17. Lang, N. D. Spectroscopy of single atoms in the scanning tunneling microscope. *Phys. Rev. B* **34**, 5947–5950 (1986).
18. Everson, M. P., Jaklevic, R. C. & Shen, W. Measurement of the local density of states on a metal surface: Scanning tunneling spectroscopic imaging of Au(111). *J. Vac. Sci. Technol. A* **8**, 3662–3665 (1990).
19. Kittel, C. Indirect exchange interactions in metals. *Solid State Phys.* **22**, 1–26 (1968).

Acknowledgements

We thank B. A. Jones, E. J. Heller, J. S. Hersch, G. Fiete, A. J. Heinrich and C. T. Rettner for helpful discussions, and L. Folks for expert assistance with image preparation.

Correspondence and requests for materials should be addressed to H.C.M. (e-mail: hari@alumni.princeton.edu).

Experimental test of quantum nonlocality in three-photon Greenberger–Horne–Zeilinger entanglement

Jian-Wei Pan*, Dik Bouwmeester†, Matthew Daniell*, Harald Weinfurter‡ & Anton Zeilinger*

* Institut für Experimentalphysik, Universität Wien, Boltzmanngasse 5, 1090 Wien, Austria

† Clarendon Laboratory, University of Oxford, Parks Road, Oxford OX1 3PU, UK

‡ Sektion Physik, Ludwig-Maximilians-Universität of München, Schellingstrasse 4/III, D-80799 München, Germany

Bell’s theorem¹ states that certain statistical correlations predicted by quantum physics for measurements on two-particle systems cannot be understood within a realistic picture based on local properties of each individual particle—even if the two particles are separated by large distances. Einstein, Podolsky and Rosen first recognized² the fundamental significance of these quantum correlations (termed ‘entanglement’ by Schrödinger³) and the

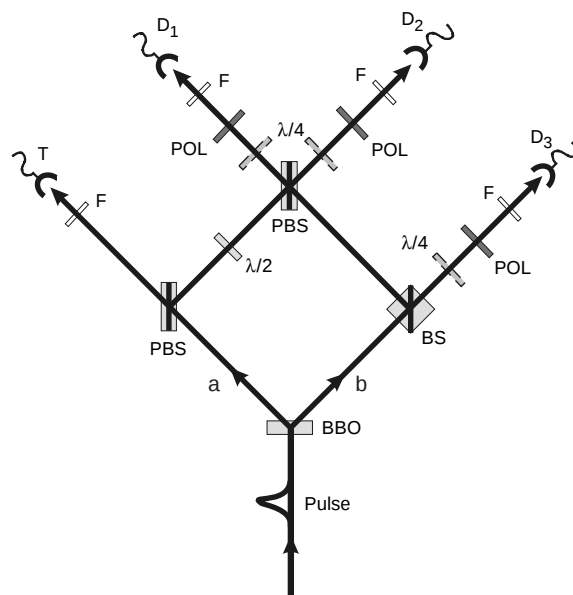


Figure 1 Experimental set-up for Greenberger–Horne–Zeilinger (GHZ) tests of quantum nonlocality. Pairs of polarization-entangled photons²⁸ (one photon H polarized and the other V) are generated by a short pulse of ultraviolet light (~ 200 fs, $\lambda = 394$ nm). Observation of the desired GHZ correlations requires fourfold coincidence and therefore two pairs²⁹. The photon registered at T is always H and thus its partner in b must be V . The photon reflected at the polarizing beam-splitter (PBS) in arm a is always V , being turned into equal superposition of V and H by the $\lambda/2$ plate, and its partner in arm b must be H . Thus if all four detectors register at the same time, the two photons in D_1 and D_2 must either both have been VV and reflected by the last PBS or HH and transmitted. The photon at D_3 was therefore H or V , respectively. Both possibilities are made indistinguishable by having equal path lengths via a and b to D_1 (D_2) and by using narrow bandwidth filters ($F \approx 4$ nm) to stretch the coherence time to about 500 fs, substantially larger than the pulse length³⁰. This effectively erases the prior correlation information and, owing to indistinguishability, the three photons registered at D_1 , D_2 and D_3 exhibit the desired GHZ correlations predicted by the state of equation (1), where for simplicity we assume the polarizations at D_3 to be defined at right angles relative to the others. Polarizers oriented at 45° and $\lambda/4$ plates in front of the detectors allow measurement of linear H/V (circular R/L) polarization.

two-particle quantum predictions have found ever-increasing experimental support⁴. A more striking conflict between quantum mechanical and local realistic predictions (for perfect correlations) has been discovered^{5,6}; but experimental verification has been difficult, as it requires entanglement between at least three particles. Here we report experimental confirmation of this conflict, using our recently developed method⁷ to observe three-photon entanglement, or 'Greenberger–Horne–Zeilinger' (GHZ) states. The results of three specific experiments, involving measurements of polarization correlations between three photons, lead to predictions for a fourth experiment; quantum physical predictions are mutually contradictory with expectations based on local realism. We find the results of the fourth experiment to be in agreement with the quantum prediction and in striking conflict with local realism.

We first analyse certain quantum predictions for the entangled three-photon GHZ state:

$$|\Psi\rangle = \frac{1}{\sqrt{2}}(|H\rangle_1|H\rangle_2|H\rangle_3 + |V\rangle_1|V\rangle_2|V\rangle_3) \quad (1)$$

where H and V denote horizontal and vertical linear polarizations respectively. This state indicates that the three photons are in a quantum superposition of the state $|H\rangle_1|H\rangle_2|H\rangle_3$ (all three are horizontally polarized) and the state $|V\rangle_1|V\rangle_2|V\rangle_3$ (all three are vertically polarized) with none of the photons having a well-defined state on its own.

We consider now measurements of linear polarization along directions H'/V' rotated by 45° with respect to the original H/V directions, or of circular polarization L/R (left-handed, right-handed). These new polarizations can be expressed in terms of the original ones as:

$$|H'\rangle = \frac{1}{\sqrt{2}}(|H\rangle + |V\rangle) \quad (2)$$

$$|V'\rangle = \frac{1}{\sqrt{2}}(|H\rangle - |V\rangle)$$

$$|R\rangle = \frac{1}{\sqrt{2}}(|H\rangle + i|V\rangle) \quad (3)$$

$$|L\rangle = \frac{1}{\sqrt{2}}(|H\rangle - i|V\rangle)$$

For convenience we will refer to a measurement of H'/V' linear polarization as an x measurement and one of R/L circular polarization as a y measurement.

Representing the GHZ state (equation (1)) in the new states by using equations (2) and (3), one obtains the quantum predictions for measurements of these new polarizations. For example, for the case of measurement of circular polarization on, say, both photon 1 and 2, and linear polarization H'/V' on photon 3, denoted as a yyx experiment, the state may be expressed as:

$$|\Psi\rangle = \frac{1}{2}(|R\rangle_1|R\rangle_2|H'\rangle_3 + |L\rangle_1|R\rangle_2|H'\rangle_3 + |R\rangle_1|R\rangle_2|V'\rangle_3 + |L\rangle_1|R\rangle_2|V'\rangle_3) \quad (4)$$

This expression implies, first, that any specific result obtained in any individual or in any two-photon joint measurement is maximally random. For example, photon 1 will exhibit polarization R or L with the same probability of 50%, or photons 1 and 2 will exhibit polarizations RL , LR , RR or LL with the same probability of 25%. Second, given any two results of measurements on any two photons, we can predict with certainty the result of the corresponding measurement performed on the third photon. For example, suppose photons 1 and 2 both exhibit right-handed R circular polar-

ization. Then by the third term in equation (4), photon 3 will definitely be V' polarized.

By cyclic permutation, we can obtain analogous expressions for any experiment measuring circular polarization on two photons and H'/V' linear polarization on the remaining one. Thus, in every one of the three yyx , xyy , and xxy experiments, any individual measurement result—both for circular polarization and for linear H'/V' polarization—can be predicted with certainty for every photon given the corresponding measurement results of the other two.

Now we will analyse the implications for local realism. As these predictions are independent both of the spatial separation and of the relative time order of the three measurements, we consider them performed simultaneously in a given reference frame—say, for conceptual simplicity, in the reference frame of the source. Then, as Einstein locality implies that no information can travel faster than the speed of light, this requires any specific measurement result obtained for any photon never to depend on which specific measurements are performed simultaneously on the other two nor on their outcome. The only way then for local realism to

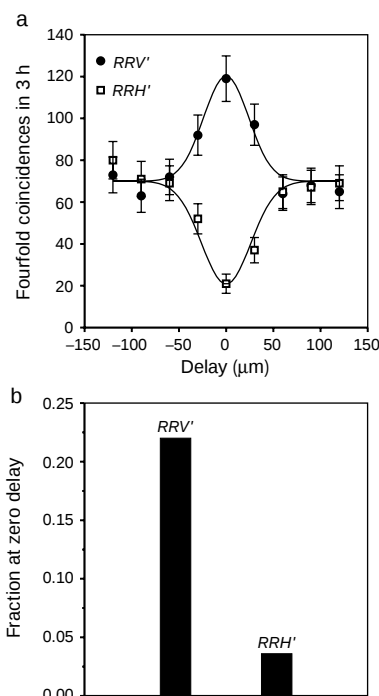


Figure 2 A typical experimental result used in the GHZ argument. This is the yyx experiment measuring circular polarization on photons 1 and 2 and linear polarization on the third. **a**, Fourfold coincidences between the trigger detector T , detectors D_1 and D_2 (both set to measure a right-handed polarized photon), and detector D_3 (set to measure a linearly polarized H' (lower curve) and V' (upper curve) photon) as a function of the delay between photon 1 and 2 at the final polarizing beam-splitter. We could adjust the time delay between paths **a** and **b** in Fig. 1 by translating the final polarizing beam-splitter (PBS) and using additional mirrors (not shown in Fig. 1) to ensure overlap of both beams, independent of mirror displacement. At large delay, that is, outside the region of coherent superposition, the two possibilities HHH and VVV are distinguishable and no entanglement results. In agreement with this explanation, it was observed within the experimental accuracy that for large delay the eight possible outcomes in the yyx experiment (and also the other experiments) have the same coincidence rate, whose mean value was chosen as a normalization standard. **b**, At zero delay maximum GHZ entanglement results; the experimentally determined fractions of RRV' and RRH' triples (out of the eight possible outcomes in the yyx experiment) are deduced from the measurements at zero delay. The fractions were obtained by dividing the normalized fourfold coincidences of a specific outcome by the sum of all possible outcomes in each experiment—here, the yyx experiment.

explain the perfect correlations predicted by equation (4) is to assume that each photon carries elements of reality for both x and y measurements that determine the specific individual measurement result^{5,6,8}.

For photon i we call these elements of reality X_i with values $+1(-1)$ for $H'(V')$ polarizations and Y_i with values $+1(-1)$ for $R(L)$; we thus obtain the relations⁸ $Y_1Y_2X_3 = -1$, $Y_1X_2Y_3 = -1$ and $X_1Y_2Y_3 = -1$, in order to be able to reproduce the quantum predictions of equation (4) and its permutations.

We now consider a fourth experiment measuring linear H'/V' polarization on all three photons, that is, an xxx experiment. We investigate the possible outcomes that will be predicted by local realism based on the elements of reality introduced to explain the earlier yyx , $yxxy$ and xyy experiments.

Because of Einstein locality any specific measurement for x must be independent of whether an x or y measurement is performed on the other photon. As $Y_iY_i = +1$, we can write $X_1X_2X_3 = (X_1Y_2Y_3)(Y_1X_2Y_3)(Y_1Y_2X_3)$ and obtain $X_1X_2X_3 = -1$. Thus from a local realist point of view the only possible results for an xxx experiment are $V'V'V'$, $H'H'V'$, $H'V'H'$, and $V'H'H'$.

How do these predictions of local realism for an xxx experiment compare with those of quantum mechanics? If we express the state given in equation (1) in terms of H'/V' polarization using equation (2) we obtain:

$$|\Psi\rangle = \frac{1}{2}(|H'\rangle_1|H'\rangle_2|H'\rangle_3 + |H'\rangle_1|V'\rangle_2|V'\rangle_3 + |V'\rangle_1|H'\rangle_2|V'\rangle_3 + |V'\rangle_1|V'\rangle_2|H'\rangle_3) \quad (5)$$

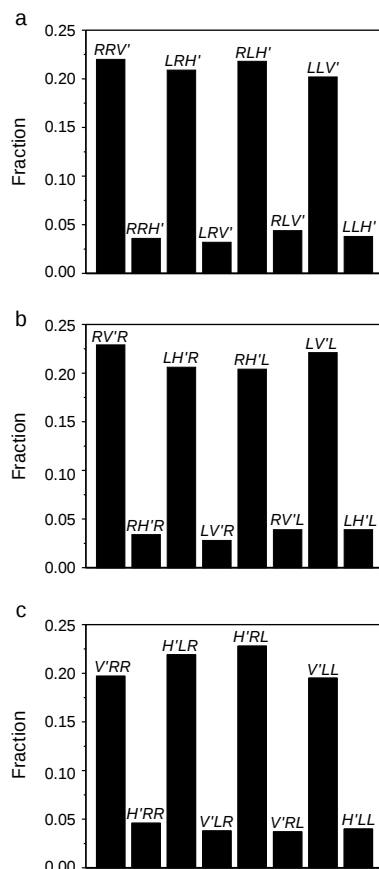


Figure 3 All outcomes observed in the yyx , $yxxy$ and xyy experiments, obtained as in Fig. 2. **a**, yyx ; **b**, $yxxy$; **c**, xyy . The experimental data show that we observe the GHZ terms predicted by quantum physics (tall bars) in a fraction of 0.85 ± 0.04 of all cases and 0.15 ± 0.02 of the spurious events (short bars) in a fraction of all cases. Within our experimental error we thus confirm the GHZ predictions for the experiments.

Thus we conclude that the local realistic model predicts none of the terms occurring in the quantum prediction and vice versa. This means that whenever local realism predicts that a specific result will definitely occur for a measurement on one of the photons based on the results for the other two, quantum physics definitely predicts the opposite result. For example, if two photons are both found to be H' polarized, local realism predicts the third photon to carry polarization V' while quantum physics predicts H' . This is the GHZ contradiction between local realism and quantum physics.

In the case of Bell's inequalities for two photons, the conflict between local realism and quantum physics arises for statistical predictions of the theory; but for three entangled particles the conflict arises even for the definite predictions. Statistics now only results from the inevitable experimental limitations occurring in any and every experiment, even in classical physics.

A diagram of our experimental set-up is given in Fig. 1. The method to produce GHZ entanglement for three spatially separated photons is a further development of the techniques that have been used in our previous experiments on quantum teleportation⁹ and entanglement swapping¹⁰. GHZ entanglement has also been inferred for nuclear spins within single molecules from NMR measurements¹¹, though there a test of nonlocality is impossible.

In the experiment GHZ entanglement is observed under the condition that the trigger detector T and the three GHZ detectors D_1 , D_2 and D_3 all actually register a photon. As there are other detection events where fewer detectors fire, this condition might raise doubts about whether such a source can be used to test local realism. The same question arose earlier for certain experiments involving photon pairs^{12,13} where a violation of Bell's inequality was only achieved under the condition that both detectors used register a photon. It was often believed^{14,15} that such experiments could never, not even in their idealized versions, be genuine tests of local realism. However, this has been disproved¹⁶. Following the same line of reasoning, it has recently been shown¹⁷ that our procedure permits a valid GHZ test of local realism. In essence, both the Bell and the GHZ arguments exhibit a conflict between detection events and the ideas of local realism.

As explained above, demonstration of the conflict between local realism and quantum mechanics for GHZ entanglement consists of four experiments, each with three spatially separated polarization measurements. First, we perform yyx , $yxxy$ and xyy experiments. If the results obtained are in agreement with the predictions for a GHZ state, then for an xxx experiment, our consequent expectations using a local-realist theory are exactly the opposite of our expectations using quantum physics.

For each experiment we have eight possible outcomes of which ideally four should never occur. Obviously, no experiment either in classical physics or in quantum mechanics can ever be perfect; therefore, even the outcomes which should not occur will occur with some small probability in any real experiment. The question is how to deal with such spurious events in view of the fact that the original GHZ argument is based on perfect correlations.

We follow two independent possible strategies. In the first strategy we simply compare our experimental results with the predictions both of quantum mechanics and of a local realist theory for GHZ correlations, and assume that the spurious events are attributable to experimental imperfection that is not correlated to the elements of reality a photon carries. A local realist might argue against that approach and suggest that the non-perfect detection events indicate that the GHZ argument is inapplicable. In our second strategy we therefore accommodate local-realist theories, by assuming that the non-perfect events in the first three experiments indicate a set of elements of reality which are in conflict with quantum mechanics. We then compare the local realist prediction for the xxx experiment obtained under that assumption with the experimental results.

The observed results for two possible outcomes in a yyx experiment

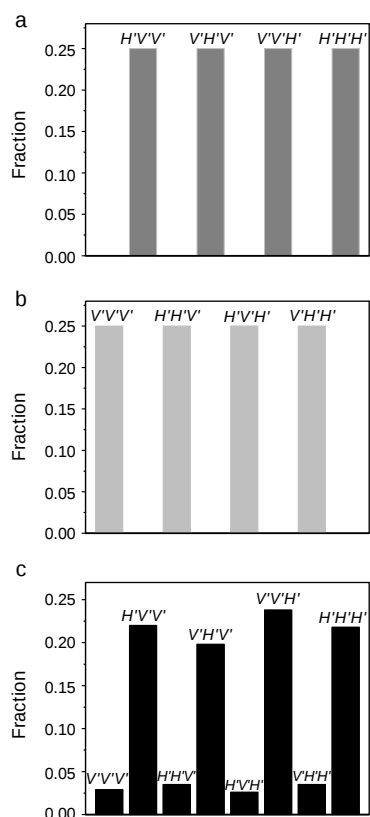


Figure 4 Predictions of quantum mechanics and of local realism, and observed results for the xxx experiment. **a**, **b**, The maximum possible conflict arises between the predictions for quantum mechanics (**a**) and local realism (**b**) because the predicted correlations are exactly opposite. **c**, The experimental results clearly confirm the quantum predictions within experimental error and are in conflict with local realism.

are shown in Fig. 2. The six remaining possible outcomes of a $yx\bar{x}$ experiment have also been measured in the same way and likewise in the $xy\bar{x}$ and $x\bar{y}y$ experiments. For all three experiments this results in 24 possible outcomes whose individual fractions thus obtained are shown in Fig. 3.

Adopting our first strategy, we assume that the spurious events are attributable to unavoidable experimental errors; within the experimental accuracy, we conclude that the desired correlations in these experiments confirm the quantum predictions for GHZ entanglement. Thus we compare the predictions of quantum mechanics and local realism with the results of an xxx experiment (Fig. 4) and we observe that, again within experimental error, the triple coincidences predicted by quantum mechanics occur and not those predicted by local realism. In this sense, we believe that we have experimentally realized the first three-particle test of local realism following the GHZ argument.

We then investigated whether local realism could reproduce the xxx experimental results shown in Fig. 4c, if we assume that the spurious non-GHZ events in the other three experiments (Fig. 3) actually indicate a deviation from quantum physics. To answer this we adopt our second strategy and consider the best prediction a local realistic theory could obtain using these spurious terms. How, for example, could a local realist obtain the quantum prediction $H'H'H'$? One possibility is to assume that triple events producing $H'H'H'$ would be described by a specific set of local hidden variables such that they would give events that are in agreement with quantum theory in both an $xy\bar{x}$ and a $y\bar{x}y$ experiment (for example, the results $H'LR$ and $LH'R$), but give a spurious event for a $yx\bar{x}$ experiment (in this case, LLH'). In this way any local realistic

prediction for an event predicted by quantum theory in our xxx experiment will use at least one spurious event in the earlier measurements together with two correct ones. Therefore, the fraction of correct events in the xxx experiment can at most be equal to the sum of the fractions of all spurious events in the $yx\bar{x}$, $y\bar{x}y$, and $x\bar{y}y$ experiments, that is, 0.45 ± 0.03 . However, we experimentally observed such terms with a fraction of 0.87 ± 0.04 (Fig. 4c), which violates the local realistic expectation by more than eight standard deviations.

Our latter argument is equivalent to adopting an inequality of the kind first proposed by Mermin¹⁸. This second analysis succeeds because our average visibility of $71\% \pm 4\%$ clearly surpasses the minimum of 50% necessary for a violation of local realism^{18–20}. However, we realise that, as for all existing two-particle tests of local realism, our experiment has rather low detection efficiencies. Therefore we had to invoke the fair sampling hypothesis^{21,22}, where it is assumed that the registered events are a faithful representative of the whole.

Possible future experiments could include: further study GHZ correlations over large distances with space-like separated randomly switched measurements²³; extending the techniques used here to the observation of multi-photon entanglement²⁴; observation of GHZ-correlations in massive objects like atoms²⁵; and investigation of possible applications in quantum computation and quantum communication protocols^{26,27}. □

1. Bell, J. S. On the Einstein-Podolsky-Rosen paradox. *Physics* **1**, 195–200 (1964); reprinted Bell, J. S. *Speakable and Unspeakeable in Quantum Mechanics* (Cambridge Univ. Press, Cambridge, 1987).
2. Einstein, A., Podolsky, B. & Rosen, N. Can quantum-mechanical description of physical reality be considered complete? *Phys. Rev.* **47**, 777–780 (1935).
3. Schrödinger, E. Die gegenwärtige Situation in der Quantenmechanik. *Naturwissenschaften* **23**, 807–812; 823–828; 844–849 (1935).
4. Aspect, A. Bell's inequality test: more ideal than ever. *Nature* **390**, 189–190 (1999).
5. Greenberger, D. M., Horne, M. A. & Zeilinger, A. in *Bell's Theorem, Quantum Theory, and Conceptions of the Universe* (ed. Kafatos, M.) 73–76 (Kluwer Academic, Dordrecht, 1989).
6. Greenberger, D. M., Horne, M. A., Shimony, A. & Zeilinger, A. Bell's theorem without inequalities. *Am. J. Phys.* **58**, 1131–1143 (1990).
7. Bouwmeester, D., Pan, J.-W., Daniell, M., Weinfurter, H. & Zeilinger, A. Observation of three-photon Greenberger-Horne-Zeilinger entanglement. *Phys. Rev. Lett.* **82**, 1345–1349 (1999).
8. Mermin, N. D. What's wrong with these elements of reality? *Phys. Today* **43**, 9–11 (1990).
9. Bouwmeester, D. *et al.* Experimental quantum teleportation. *Nature* **390**, 575–579 (1997).
10. Pan, J.-W., Bouwmeester, D., Weinfurter, H. & Zeilinger, A. Experimental entanglement swapping: Entangling photons that never interacted. *Phys. Rev. Lett.* **80**, 3891–3894 (1998).
11. Laflamme, R., Knill, E., Zurek, W. H., Catasti, P. & Mariappan, S. V. S. NMR Greenberger-Horne-Zeilinger states. *Phil. Trans. R. Soc. Lond. A* **356**, 1941–1947 (1998).
12. Ou, Z. Y. & Mandel, L. Violation of Bell's inequality and classical probability in a two-photon correlation experiment. *Phys. Rev. Lett.* **61**, 50–53 (1988).
13. Shih, Y. H. & Alley, C. O. New type of Einstein-Podolsky-Rosen-Bohm experiment using pairs of light quanta produced by optical parametric down conversion. *Phys. Rev. Lett.* **61**, 2921–2924 (1988).
14. Kwiat, P., Eberhard, P. E., Steinberger, A. M. & Chiao, R. Y. Proposal for a loophole-free Bell inequality experiment. *Phys. Rev. A* **49**, 3209–3220 (1994).
15. De Caro, L. & Garuccio, A. Reliability of Bell-inequality measurements using polarization correlations in parametric-down-conversion photons. *Phys. Rev. A* **50**, R2803–R2805 (1994).
16. Popescu, S., Hardy, L. & Zukowski, M. Revisiting Bell's theorem for a class of down-conversion experiments. *Phys. Rev. A* **56**, R4353–R4357 (1997).
17. Zukowski, M. Violations of local realism in multiphoton interference experiments. *Phys. Rev. A* **61**, 022109 (2000).
18. Mermin, N. D. Extreme quantum entanglement in a superposition of macroscopically distinct states. *Phys. Rev. Lett.* **65**, 1838–1841 (1990).
19. Roy, S. M. & Singh, V. Tests of signal locality and Einstein-Bell locality for multiparticle systems. *Phys. Rev. Lett.* **67**, 2761–2764 (1991).
20. Zukowski, M. & Kaszlikowski, D. Critical visibility for N -particle Greenberger-Horne-Zeilinger correlations to violate local realism. *Phys. Rev. A* **56**, R1682–R1685 (1997).
21. Pearle, P. Hidden-variable example based upon data rejection. *Phys. Rev. D* **2**, 1418–1425 (1970).
22. Clauser, J. & Shimony, A. Bell's theorem: experimental tests and implications. *Rep. Prog. Phys.* **41**, 1881–1927 (1978).
23. Weihs, G., Jennewein, T., Simon, C., Weinfurter, H. & Zeilinger, A. Violation of Bell's inequality under strict Einstein locality conditions. *Phys. Rev. Lett.* **81**, 5039–5043 (1998).
24. Bose, S., Vedral, V. & Knight, P. L. Multiparticle generalization of entanglement swapping. *Phys. Rev. A* **57**, 822–829 (1998).
25. Haroche, S. Atoms and photons in high-Q cavities: next tests of quantum theory. *Ann. NY Acad. Sci.* **755**, 73–86, (1995).
26. Briegel, H.-J., Duer, W., Cirac, J. I. & Zoller, P. Quantum repeaters: The role of imperfect local operations in quantum communication. *Phys. Rev. Lett.* **81**, 5932–5935 (1998).
27. Cleve, R. & Buhrman, H. Substituting quantum entanglement for communication. *Phys. Rev. A* **56**, 1201–1204 (1997).
28. Kwiat, P. G. *et al.* New high intensity source of polarization-entangled photon pairs. *Phys. Rev. Lett.* **75**, 4337–4341 (1995).

29. Zeilinger, A., Horne, M. A., Weinfurter, H. & Zukowski, M. Three particle entanglements from two entangled pairs. *Phys. Rev. Lett.* **78**, 3031–3034 (1997).
30. Zukowski, M., Zeilinger, A. & Weinfurter, H. Entangling photons radiated by independent pulsed source. *Ann. NY Acad. Sci.* **755**, 91–102 (1995).

Acknowledgements

We thank D. M. Greenberger, M. A. Horne and M. Zukowski for useful discussions. This work was supported by the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung, the Austrian Academy of Sciences and the Training and Mobility of Researchers programme of the European Union.

Correspondence and requests for materials should be addressed to A.Z. (e-mail: Zeilinger-office@exp.univie.ac.at).

Ball lightning caused by oxidation of nanoparticle networks from normal lightning strikes on soil

John Abrahamson & James Dinniss

Chemical and Process Engineering Department, University of Canterbury, Private Bag 4800, Christchurch, New Zealand

Observations of ball lightning have been reported for centuries, but the origin of this phenomenon remains an enigma. The 'average' ball lightning appears as a sphere with a diameter of 300 mm, a lifetime of about 10 s, and a luminosity similar to a 100-W lamp¹. It floats freely in the air, and ends either in an explosion, or by simply fading from view. It almost invariably occurs during stormy weather^{2,3}. Several energy sources have been proposed^{2–4} to explain the light, but none of these models has succeeded in explaining all of the observed characteristics. Here we report a model that potentially accounts for all of those properties, and which has some experimental support. When normal lightning strikes soil, chemical energy is stored in nanoparticles of Si, SiO or SiC, which are ejected into the air as a filamentary network. As the particles are slowly oxidized in air, the stored energy is released as heat and light. We investigated this basic process by exposing soil samples to a lightning-like discharge, which produced chain aggregates of nanoparticles: these particles oxidize at a rate appropriate for explaining the lifetime of ball lightning.

Away from buildings, the material most commonly in the path of a lightning strike is a tree, and then soil. Lightning leaves solid tubular or lumpy residues (fulgurites) after interacting with sand or soil, which indicate that the discharge has penetrated beneath the ground surface, and that the material has been molten. If soils or tree roots are regarded as a fine mixture of silica and carbon, then under such high-temperature treatment, one expects chemical reduction to silicon metal, silicon monoxide, or silicon carbide, followed by oxidation by oxygen from the air.

Such reduction of a C/SiO₂ mix using an electric arc is commonly used in industry. Liquid silicon is the dominant equilibrium condensed phase around 3,000 K (ref. 5) for C/SiO₂ mole ratios of 1–2, with solid SiC expected for ratios >2. This ratio can range from 0.1 to about 2 for mineral soils, and is much higher for wood. Silicon metal has been observed⁶ in the silicate glass deposit adjacent to a charred tree root after a large lightning strike.

Such fast-cooling processes often yield nanometre-sized particles^{7,8}. For lightning action on a soil, or a soil/wood mix with a C/SiO₂ ratio of 1–2, we expect the particles to be Si and SiO

(formed by condensation of the dominant vapour species⁵), with SiC and soot dominating for C/SiO₂ > 2. Most nanoparticle suspensions are found in the form of chain aggregates⁹, which extend where charged particles influence their neighbours¹⁰ (in conditions of higher particle charge and numbers, and fewer gaseous ions, as at cooler flame temperatures). Charge on the growing chain may induce a dendrimer-like structure, growing from the centre. The possible size is suggested by the following observations in quiescent air. Filamentary particle structures spanning 50 mm have been found¹¹ after vaporizing metal in air in the presence of electric fields. Early work with charged aerosols¹² showed the formation of a spherical networked aerosol suspension of diameter 200 mm.

Particle networks have been proposed more recently¹³ as a general basis for ball lightning, with the aggregation influenced by the field of the growing ball¹, but without a clear proposal for the chemical reaction occurring. Oxidation of copper particles has also been suggested¹⁴, but with this process occurring at the perimeter of the sphere of air carrying the particles: a network structure or rate limitation at the particle surface was not mentioned.

A Si/SiO/SiC nanoparticle network would have a large surface, and could be expected to oxidize rapidly, even explosively. However, we emphasize that the rate of oxidation would be limited by the need for oxygen to diffuse through the developing SiO₂ layer to the metal (or carbide) beneath. Laboratory oxidation studies on silicon surfaces^{15,16} show that both oxygen and water are reactants. Whether oxygen or water dominate the reaction with silicon depends on their partial pressures. SiC oxidizes at a similar rate to Si (ref. 17).

We checked for the existence of nanosphere chains after exposing soil to a lightning-like discharge. A 10–20 kV d.c. discharge penetrated a 3-mm layer of soil, transferring up to 3.4 C of charge. Sampling of the air space close to the discharge caused deposits to form on a glass-fibre filter and a filter-mounted transmission electron microscope (TEM) grid. Examination of these deposits using scanning electron microscopy (SEM) showed 'lumpy' filament links (width 100 nm, length 10 µm) between the glass fibres. TEM at high resolution (Fig. 1) showed chain aggregates of nanospheres, 5–70 nm in diameter. Larger spheres, several micrometres in diameter, were collected on all the filters, and were collectively of similar mass to the nanoparticles.

Despite using charge transfer within the range observed for lightning strikes¹⁸, we did not observe any luminous ball. At the higher power levels, the soil sample was always completely blown away in the radial direction. It appeared unlikely that a network of delicate long filaments could survive the discharge shockwave. If the

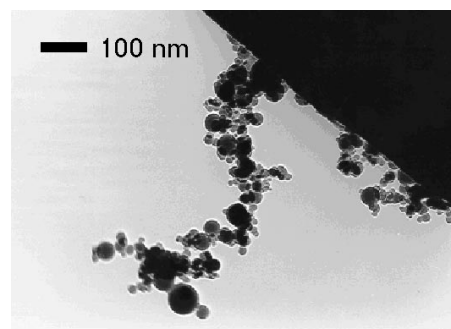


Figure 1 Transmission electron micrograph of nanoparticle chains sampled from the discharge environment. These particles were deposited on a nickel grid, after sampling the gas space above a 14.9-kV discharge on a silt loam soil containing 12.5% carbon. The soil was placed in a layer on a flat conducting (graphite) base below a vertical graphite electrode, with a gap of 22 mm to the soil, and 3.0 C charge transferred from a 204 µF capacitor. The extended chains are made up of spheres 5–70 nm in diameter; the chain width is 25–125 nm. Six of the highest-power runs were examined in this way, with three soils. All showed similar results, except one showing only larger particles.

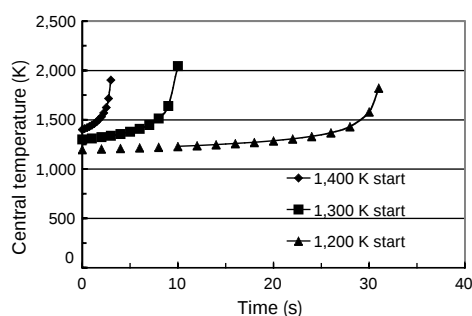


Figure 2 Temperature history of a 300-mm-diameter ball lightning predicted from our nanoparticle network model. The temperature of the central region is given for starting temperatures 1,200, 1,300 and 1,400 K. The ball has 100 g m^{-3} Si loading in 25-nm spheres. Lifetimes before melting are around 30, 10 and 3 s, respectively. For a temperature of 1,473 K, the rate of heat release is estimated at 79.0 kW m^{-2} of ball volume. The rate has been corrected for the partial pressures of O_2 and H_2O , yielding 0.10 nm s^{-1} oxide growth at 1,473 K for either reactant. Estimation for other temperatures uses an activation energy¹⁵ of 189 kJ mol^{-1} .

observed filaments form the basis of ball lightning, they must somehow avoid this shockwave. Lightning strikes to ground penetrate a deeper soil layer than in our experiments. Radial dispersal of vapour/hot particles is then curtailed by the surrounding soil. This hot material from the fulgurite cavity would instead disperse up into the air, after decay of the strike pressure.

This action of a fulgurite cavity has been simulated¹⁹ by an evacuated chamber with an opening to the atmosphere temporarily sealed with a breakable diaphragm. A high voltage discharge acted on the chamber wall materials (ice and plastics), breaking the diaphragm. A luminous spherical ball (150–400 mm dia.) was ejected that persisted for several milliseconds (ref. 19). These observations showed all the features of a “smoke-ring vortex”, the toroidal vortex formed by sudden jetting of a fluid through a circular opening²⁰. The vortex above a fulgurite cavity is expected to form its own spherical quiescent space, thereby allowing formation of delicate filaments. The materials chosen above¹⁹ did not contain metal elements, eliminating any long-lived oxidation and luminosity. Soot nanoparticles from the plastic pyrolysis would oxidize in several milliseconds.

Ball lightning is commonly spherical, and shows elastic behaviour. This is consistent with observed elastic properties of nanoparticle chains⁹. Also, the repulsion of like particle charges will keep the net from collapsing in on itself, with those at the ball surface giving a surface-tension effect to the ball¹, restoring the spherical shape after deformation against solid boundaries. The ball will generally follow the wind but electrical forces (caused by interactions with, for example, nearby conductors) may also be important.

Our quantitative model for a net-filled ball lightning assumes negligible temperature difference between the network and the enclosed gas, and negligible local relative motion between network and gas. The nanoparticle Si concentration can be given a maximum value, around 100 g m^{-3} , derived from vapour Si at 1 atmosphere pressure and 3,000–5,000 K (ref. 5), cooled to around 1,000–1,500 K. The jet from the fulgurite will also pick up an uncertain amount of non-vaporized particles of soil (micrometre-sizes) from the channel sides. By mixing, these will rapidly cool the jet to a particular final temperature. If the ball that forms has neutral buoyancy (as commonly observed), the extra loading of soil is about 800 g m^{-3} . The soil particles are expected to have melted like those observed in our SEM study, and to become attached to the network. They add to the specific heat capacity of the ball. We choose a ball diameter of 300 mm (within the size range of luminous

fulgurite ejection¹⁹, equal to the mean observed size of ball lightning¹).

The transient (symmetrical) temperature profile T within the ball can be evaluated from energy conservation for a thin shell of the ball between radius r and $r + \delta r$ over a time interval δt :

$$(\rho C_p)_{\text{ball}}(4\pi r^2 \delta r) \delta T_{\text{rise}} = (-\delta(4\pi r^2 q) + H(4\pi r^2 \delta r)) \delta t \quad (1)$$

where the thermal flux q is given by $q = -k \delta T / \delta r$, and k , ρ , C_p are thermal conductivity, density and specific heat of the ball; H is the heat evolution per unit volume. Equation (1) was integrated using spreadsheet software, with property values using the assumptions below. The temperature profiles at various times are flat, except for a sharp drop near the surface. Thus the interior of the ball is not affected by heat losses from the surface over the likely life of the ball. Figure 2 shows the central plateau temperature of a ball with diameter $D = 300 \text{ mm}$, formed from 25-nm spheres, as a function of time, for three starting temperatures. The exponential rise of H with temperature yields abrupt increases after around 3–30 s for starting temperatures of 1,400–1,200 K. Here the model predicts a violent end to the ball. As the temperature rises above 1,700 or 2,000 K, silicon or silica will melt, allowing faster oxidation and network break-up. But for lower reactive-metal loadings (for example, ball lightning from low-carbon soils), the melting temperature may not be reached, with the ball fading from view. For lower starting temperatures, the ball may become visible only over the later part of its lifetime, which could be after minutes, so that it is not easily connected with the formative lightning strike.

Radiation transfer is assumed to not significantly affect the energy balance of the ball (discussed below). The k of the ball is taken to be that of air, ignoring the solids contribution (there is little connected cross-section). Energy loss from the exterior is by natural convection. Energy dissipation H increases as the surface area of the nanospheres, using the early (constant) rate per m^2 for planar surfaces of silicon¹⁵. Any soot has been already oxidized shortly after ball formation. Nanoparticle SiO or SiC are not considered in these numerical results, but will oxidize like Si (if SiO, with a lower heat of oxidation²¹). Finally, H is not limited by diffusion of oxygen into the ball; that is, the mole fraction of oxygen gas remains constant at all ball radii (an excellent approximation at $<1,400 \text{ K}$).

The above assumptions give rates of external energy loss by convection of 20–40 W for a 300-mm ball once the inner temperature reaches around 1,450 K. The oxidation energy release rate is much higher (here 1,100 W), and is largely absorbed by heating the ball mass. The total chemical energy density is 3.3 MJ m^{-3} , lying within the very approximate observational estimates of 1.5–15 MJ m^{-3} for average ball lightning¹. A cool outer surface and lack of impression of (radiant) heat are natural outcomes of this model, corresponding to generally observed properties^{2,3}.

We now consider the effective luminosity. The simplest assumption is that the nanospheres emit or absorb according to optical properties of bulk materials. This is adequate for particles $>10 \text{ nm}$ diameter²². Rayleigh’s approximation²³ for the absorption cross-section of a sphere allows one to estimate the linear absorption coefficient, α_{abs} , of the nanosphere suspension for various wavelengths, using listed values of extinction coefficient κ and refractive index n for crystalline Si, crystalline SiO₂ and amorphous SiO (ref. 24). For 100 g m^{-3} Si in the form of Si, SiO or SiO₂, the ball is optically transparent in the visible range (for Si at a wavelength of 600 nm, $\alpha_{\text{abs}} = 0.26 \text{ m}^{-1}$, so $\alpha_{\text{abs}} D = 0.08 \ll 1$), allowing one to sum the contributions of all nanoparticles for the ball emission. Emissivity is taken as equal to absorptivity.

The emission is the product of emissivity and the blackbody emission e_b for each wavelength λ ; this gives a sharp peak in the red part of the spectrum for both Si and SiO, because $d\kappa/d\lambda$ is negative and counters the positive $de_b/d\lambda$. At 1,200–1,400 K, 4.5–35 W is emitted from the whole Si peak, corresponding to 1.2–14 W over the

visible range 400–800 nm. A 100-W tungsten filament lamp (observed distribution median¹) emits 8 W in the visible (with its filament at 3,000 K), matching this nanosphere model at 1,350 K. The λ dependence of Si emission at 1,200 K matches a 1,700 K blackbody profile in the visible, so such a ball would be reported as ‘translucent white’. Lower temperatures may be reported as yellow or red, thus covering most of the observed colours. Optical emission may also come from molecules of salts vaporized from the soil. Many soils are expected to yield SiO rather than Si. Amorphous Si may also be formed, with a higher absorptivity than if crystalline²⁵. (We note that some lightning strikes (or man-made discharges) on metal structures may also result in a long-lived oxidation of networked metal nanospheres.)

The model of a nanoparticle network of slowly oxidizing Si that we present here successfully explains all the salient observed features of ball lightning in its most common environment. Future work may usefully explore the lightning/soil interaction, the chemical properties of the soil (or soil/wood mix) and lightning parameters (strength, duration), to find those which produce ball lightning in the laboratory. □

Received 1 September; accepted 10 November 1999.

- Smirnov, B. M. The properties and the nature of ball lightning. *Phys. Rep.* **152**, 177–226 (1987).
- Barry, J. D. *Ball Lightning and Bead Lightning* (Plenum, New York, 1980).
- Turner, D. J. The structure and stability of ball lightning. *Phil. Trans. R. Soc. Lond. A* **347**, 83–111 (1994).
- Singer, S. *The Nature of Ball Lightning* (Plenum, New York, 1971).
- Hutchison, S. G., Richardson, L. S. & Wai, C. M. Carbothermic reduction of silicon dioxide—a thermodynamic investigation. *Metall. Trans. B* **19**, 249–253 (1988).
- Essene, E. J. & Fisher, D. C. Lightning strike fusion: extreme reduction and metal-silicate liquid immiscibility. *Science* **234**, 189–193 (1986).
- Siegel, R. W. in *Physics of New Materials* 2nd edn (ed. Fujita, F. E.) 66–106 (Springer, Heidelberg, 1998).
- Ludwig, M. H. in *Handbook of Optical Properties* Vol. II, *Optics of Small Particles, Interfaces and Surfaces* (eds Hummel, R. E. & Wissmann, P.) 103–127 (CRC, Boca Raton, 1997).
- Friedlander, S. I., Jang, H. D. & Ryu, K. H. Elastic behaviour of nanoparticle chain aggregates. *Appl. Phys. Lett.* **72**, 173–175 (1998).
- Fortov, V. E. *et al.* Highly nonideal classical thermal plasmas: experimental study of ordered macroparticle structures. *JETP* **84**, 256–261 (1997).
- Aleksandrov, V. Ya., Borodin, I. P., Kechenko, E. V. & Podmoshenskii, I. V. Rapid coagulation of submicron aerosols into filamentary three-dimensional structures. *Sov. Phys. Tech. Phys.* **27**, 527–529 (1982).
- Cawood, W. & Patterson, H. S. A curious phenomenon shown by highly charged aerosols. *Nature* **128**, 150 (1931).
- Aleksandrov, V. Ya., Golubev, E. M. & Podmoshenskii, I. V. Aerosol mode of ball lightning. *Sov. Phys. Tech. Phys.* **27**, 1221–1224 (1983).
- Lowke, J. J., Uman, M. A. & Liebermann, R. W. Toward a theory of ball lightning. *J. Geophys. Res.* **74**, 6887–6898 (1969).
- Deal, B. E. & Grove, A. S. General relationship for the thermal oxidation of silicon. *J. Appl. Phys.* **36**, 3770–3778 (1963).
- Massoud, H. Z. & Plummer, J. D. Analytical relationship for the oxidation of silicon in dry oxygen in the thin-film regime. *J. Appl. Phys.* **62**, 3416–3423 (1987).
- Jacobson, N. S. Corrosion of silicon-based ceramics in combustion environments. *J. Am. Ceram. Soc.* **76**, 3–28 (1993).
- Rayle, W. D. *Ball Lightning Characteristics* (Note TND-3188, NASA, Washington, 1966).
- Andrianov, A. M. & Sinitsyn, V. I. Erosion-discharge model for ball lightning. *Sov. Phys. Tech. Phys.* **22**, 1342–1347 (1977).
- Pemberton, S. T. & Davidson, J. F. Turbulence in the freeboard of a gas-fluidized bed. The significance of ghost bubbles. *Chem. Eng. Sci.* **39**, 829–840 (1984).
- Nagamori, M., Boivin, J. A. & Claveau, A. Gibbs free energies of formation of amorphous Si₂O₃, SiO and SiO₂. *J. Non-Cryst. Solids* **189**, 270–276 (1995).
- Littau, K. A. *et al.* A luminescent silicon nanocrystal colloid via a high-temperature aerosol reaction. *J. Phys. Chem.* **97**, 1224–1230 (1993).
- Bohren, C. F. & Huffman, D. R. *Absorption and Scattering of Light by Small Particles* (Wiley, New York, 1983).
- Palik, E. D. *Handbook of Optical Constants of Solids* (Academic, Orlando, 1985).
- INSPEC *Properties of Silicon* 75, 938 (Institution of Electrical Engineers, London, 1988).

Acknowledgements

We thank the Electrical and Electronic Engineering Department at the University of Canterbury, especially J. Woudberg, for the use of their high-voltage laboratory; C. Maslin, B. Lane, P. Niamskul and T. Benson for experimental work, supporting that of J. Dinniss, and D. Brown, N. Foot and R. Boyce for technical help. We also thank N. Andrews and J. McKenzie of the Plant and Microbial Sciences Department for electron microscopy.

Correspondence and requests for materials should be addressed to J.A. (e-mail: j.abrahamson@cape.canterbury.ac.nz).

Large-scale complementary integrated circuits based on organic transistors

B. Crone*, A. Dodabalapur*, Y.-Y. Lin*, R. W. Filas*, Z. Bao*, A. LaDuca†, R. Sarpeshkar*‡, H. E. Katz* & W. Li*

* Bell Laboratories, Lucent Technologies, 600 Mountain Avenue, Murray Hill, New Jersey 07974, USA

† Lucent Technologies, 555 Union Boulevard, Allentown, Pennsylvania 18103, USA

‡ Present address: Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Thin-film transistors based on molecular and polymeric organic materials have been proposed for a number of applications, such as displays^{1–3} and radio-frequency identification tags^{4–6}. The main factors motivating investigations of organic transistors are their lower cost and simpler packaging, relative to conventional inorganic electronics, and their compatibility with flexible substrates^{7,8}. In most digital circuitry, minimal power dissipation and stability of performance against transistor parameter variations are crucial. In silicon-based microelectronics, these are achieved through the use of complementary logic—which incorporates both p- and n-type transistors—and it is therefore reasonable to suppose that adoption of such an approach with organic semiconductors will similarly result in reduced power dissipation, improved noise margins and greater operational stability. Complementary inverters and ring oscillators have already been reported^{9,10}. Here we show that such an approach can realize much larger scales of integration (in the present case, up to 864 transistors per circuit) and operation speeds of ~1 kHz in clocked sequential complementary circuits.

Our fabrication method for organic complementary circuits spatially separates n-channel and p-channel transistors to facilitate the separate deposition of electron-transporting and hole-transporting semiconductors. Other approaches are possible, and an alternative method will also be outlined. In the experiments described below, hexadecafluorocopper phthalocyanine (F-CuPc) is used as the n-type semiconductor and α -sexithiophene (α -6T) as the p-type semiconductor. The chemical structures of these

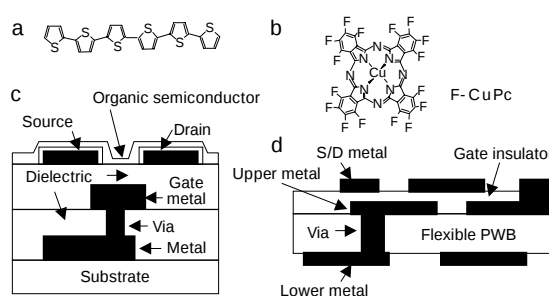


Figure 1 Chemical structures of organic semiconductors and schematic layer structures of circuits. **a**, Structure of α -sexithiophene (α -6T); **b**, structure of hexadecafluorocopper phthalocyanine (F-CuPc). **c**, Schematic layer structure of the circuits used in this study. The circuits are fabricated on rigid substrates, and metal lines and vias are defined by photolithography and standard semiconductor processing techniques. The interconnection metal level is defined immediately above the substrate, followed by the gate metal level and the source/drain (S/D) metal level. The interlayer dielectrics are not shown. The organic semiconductor is deposited above the S/D metal. **d**, Schematic of the cross-section of an organic transistor circuit technology based on flexible printed wiring boards.

materials are shown in Fig. 1a and b. Both F-CuPc (ref. 11) and α -6T (refs 12, 13) are examples of relatively well ordered organic semiconductors with mobilities in excess of $10^{-2} \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ in thin-film form. There are other possible choices of well ordered organic and polymer semiconductors, particularly for forming the p-channel transistors. Some alternatives include pentacene (ref. 3), regioregular poly(thiophene) (refs 1, 2), and oligothiophenes other than α -6T.

The vertical circuit layout that we employ is based on an 'upside-down' configuration (Fig. 1c), in which the interconnect metal level is defined first, followed by the gate metal level, and finally the source/drain metal level. The gate and source/drain metal levels are also used for local interconnections. This layout configuration differs from that employed in Si-based circuits and in the polymer circuits reported by Drury *et al.*⁵, wherein the transistors are formed first followed by the interconnect metallization. Figure 1c shows the schematic of the circuits formed on rigid substrates that we employed for this work, whereas Fig. 1d shows the schematic of a process that we have developed for circuits based on flexible plastic substrates. Standard flexible printed wiring board processes allow the fabrication of two levels of metal on either side of an insulating substrate and through-board conductors (vias) which connect these two levels. The application of a thin dielectric (which functions as the gate dielectric) above the upper metal level, together with the definition of a third metal level (source/drain, S/D, metal), results in a layer structure with a cross-sectional topology similar to that shown in Fig. 1c. This is an example of an approach to the fabrication of flexible circuits that is essentially an extension of existing technology and is both reliable and low-cost.

The operation of discrete devices as well as circuits depends critically on the metallization employed for the sources and drains. We have found that a combination of electroless Ni deposition and immersion Au coating¹⁴ works very well for both n-channel and p-channel thin-film transistors (TFTs). These metals are deposited from appropriate solutions over pre-patterned materials that can be sensitized for electroless deposition. Five-stage ring oscillators have an oscillation frequency of 10 kHz for devices with channel lengths of 7.5 μm , providing a useful index for the speeds of circuits that can be made. Although this is the highest frequency reported so far for ring oscillators based on organic semiconductors, it can be increased further by shrinking the channel length and/or employing higher-mobility semiconductors. Reducing device and stray capacitances will also be helpful in improving circuit speeds.

The design of logic circuits was based on a particular type of complementary logic known as pass transistor logic¹⁵. The essential circuit building blocks are an inverter and a transmission gate, as shown in Fig. 2. These building blocks have been combined to form latches and D flip-flops (Fig. 2e). The use of pass transistor logic

results in space and power-dissipation reduction, as fewer transistors are required to implement most functions. The relatively simple layouts of pass transistors logic are convenient for implementation on flexible plastic substrates.

Complementary circuits that we have tested include row decoders and shift registers. The largest circuit that was evaluated was a 48-stage shift register with 24 output buffers. The total number of transistors in this circuit is 864, and the channel length of each transistor is 7.5 μm . Each stage of the shift register is a D flip-flop, with the output of one stage connected to the input of the next. The clock and its complement drive all the stages. Every second stage has an output buffer, which consists of two inverters with large transistors to facilitate probing without loading the circuit. Shift registers are ubiquitous in digital systems and can perform many functions. One function is to shift a 'bit' in an orderly and predictable manner from one stage to the next every clock cycle. The operation of the shift register described above is illustrated in Fig. 3, in which the clock, data, and output of the 24 output buffers are plotted as a function of time. The data consist of a single bit, which is sequentially shifted over all stages of the register. We have operated two-stage shift registers at clock rates of up to 1 kHz. The operating voltage is 80 V because of the thick gate dielectric employed. By utilizing thinner or higher-*k* dielectrics¹⁶, the operating voltage as well as the power dissipation can be reduced.

The static current drawn by the 48-stage register, including the output buffers, is 70 μA , and that drawn by a smaller two-stage register (with 1 buffer per stage) is 7.6 μA . In comparison, two-stage complementary shift registers based on NOR/NAND gates required more than twice as many transistors as the two-stage shift register based on pass transistor logic, and drew 20 μA . Simulations indicate that the current drawn by shift registers based on p-channel transistors alone is greater than that of complementary shift registers of comparable transistor dimensions and speed. The static current drawn in p-channel TFT registers depends strongly on the nature of the load (whether enhancement or depletion), the on-off current ratio, and the ratio of the dimensions of the transistors that constitute the basic gates (NOR, NAND and inverter). The static current drawn by complementary registers can be further reduced by reducing the off-current of the transistors. The lower power dissipation of complementary circuits will be an important factor in an application such as radio-frequency identification tags, which are powered by the ambient field.

The schematic of a three-bit row decoder is shown in Fig. 4a. This decoder is designed so that a single serial input activates one of eight outputs. It consists of three D flip-flops and a NOR array (Fig. 4b). The output of the three flip-flops and their complements drive the NOR array which is configured so that for any given set of inputs only one output is 'high'. The characteristics of the decoder are

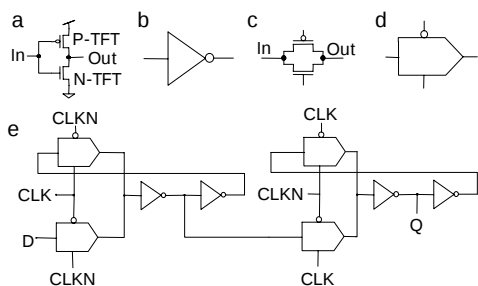


Figure 2 Representation of circuits and their constituents. **a**, A complementary inverter; **b**, symbol of an inverter; **c**, layout of a complementary transmission gate; **d**, symbol of a transmission gate; **e**, schematic of a D flip-flop showing the data input (D) and the connections of the clock (CLK) and its complement (CLKN). Also shown is the output (Q).

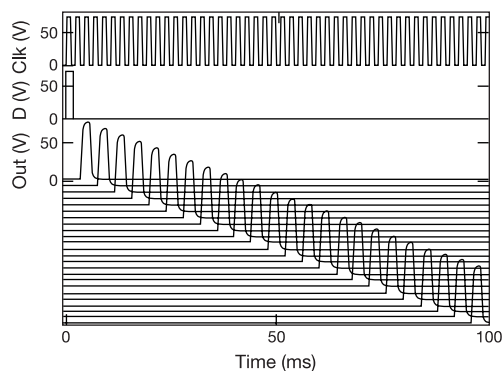


Figure 3 Characteristics of the 48-stage shift register. The clock (500 Hz) and data are shown along with the output voltages of the 24 output buffers as a function of time. The 24 outputs have been vertically offset for clarity.

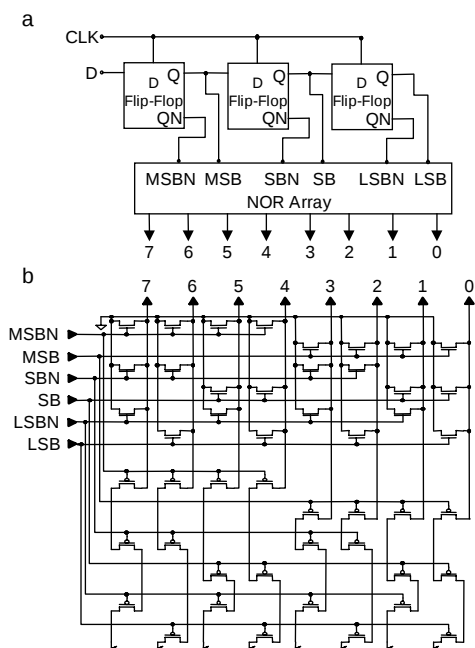


Figure 4 Row decoder design. **a**, Schematic of a three-bit row decoder with eight outputs (0–7). Each square is a D flip-flop (Fig. 2e). The bits are designated as most significant bit (MSB), second bit (SB) and least significant bit (LSB). Complements are labelled with a N. **b**, Details of the NOR array showing the manner in which individual transistors are connected.

shown in Fig. 5. The circuit used for the NOR array is significant for another reason: four transistors are serially connected between supply and ground. Such ‘four-deep’ connections are often employed in silicon arithmetic and logical unit (ALU) circuits. The fact that such a configuration has been shown to work for organic semiconductors suggests that complex logic gates such as those used in ALUs and simple microprocessors can be constructed out of organic TFTs.

There are alternative approaches to the formation of complementary circuits that differ from that described above. Instead of spatially separating the n-channel and p-channel TFTs, it is possible to create heterojunction organic transistors with two active layers, one of which is electron-transporting and the other hole-transporting¹⁷. Thus, individual devices can function either as an n-channel or a p-channel device depending on bias conditions. A hybrid organic/inorganic complementary circuit technology with inorganic and organic semiconductors has been reported^{18,19}. The propagation delay per gate in a five-stage ring oscillator with amorphous Si/organic transistors has been measured as 5 μ s (ref. 19), corresponding to an oscillation frequency of 20 kHz.

In addition to their lower static power dissipation, complementary circuits have other advantages which are important from the perspective of system lifetime. Complementary circuits are more robust against variation of transistor parameters than are p-FET (field-effect transistor) or n-FET circuits. Further, complementary circuits are expected to last longer as they draw less current. With these many advantages, we anticipate that large-scale complementary logic circuits based on organic/polymer transistors will be attractive for numerous applications. □

Received 31 August; accepted 25 November 1999.

1. Sirringhaus, H., Tessler, N. & Friend, R. H. Integrated optoelectronic devices based on conjugated polymers. *Science* **280**, 1741–1744 (1998).

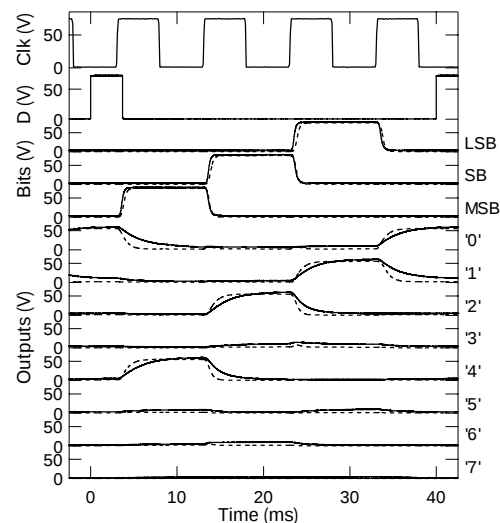


Figure 5 Characteristics of the three-bit decoder. In descending order, the traces are as follows. Clock, data, outputs of the three D flip-flops (labelled LSB, SB, MSB on the right-hand vertical axis). The eight remaining traces, labelled 0–7, show the output voltages of the eight outputs: at any time, one output at most is high. Also shown (dashed lines) are the simulated responses of the flip-flops and the outputs of the decoder. The good agreement between simulation and experiment may be noted. The simulations were performed with a set of tools developed for organic/polymer transistor circuit design.

2. Dodabalapur, A. *et al.* Organic smart pixels. *Appl. Phys. Lett.* **73**, 142–144 (1998).
3. Jackson, T. N. *et al.* Organic thin-film transistors for organic light-emitting flat-panel display backplanes. *IEEE J. Spec. Topics. Quant. Electron.* **4**, 100–104 (1998).
4. Brown, A. R. *et al.* Logic gates made from polymer transistors and their use in ring oscillators. *Science* **270**, 972–974 (1995).
5. Drury, C. *et al.* Low cost all-polymer integrated circuits. *Appl. Phys. Lett.* **73**, 108–111 (1999).
6. Ziemelis, K. Putting it on plastic. *Nature* **394**, 619–620 (1998).
7. Garnier, F. *et al.* Thin film transistors based on organic conjugated semiconductors. *Chem. Phys.* **227**, 253–262 (1996).
8. Bao, Z. *et al.* High-performance plastic transistors fabricated by printing techniques. *Chem. Mater.* **9**, 1299–1301 (1997).
9. Dodabalapur, A. *et al.* Complementary circuits with organic transistors. *Appl. Phys. Lett.* **69**, 4227–4229 (1996).
10. Lin, Y.-Y. *et al.* Organic complementary ring oscillators. *Appl. Phys. Lett.* **74**, 2714–2716 (1999).
11. Bao, Z., Lovinger, A. J. & Brown, J. New air-stable n-channel organic thin-film transistors. *J. Am. Chem. Soc.* **120**, 207–208 (1998).
12. Dodabalapur, A., Torsi, L. & Katz, H. E. Organic transistors: two dimensional transport and improved electrical characteristics. *Science* **268**, 270–271 (1995).
13. Katz, H. E., Torsi, L. & Dodabalapur, A. Synthesis, material properties and transistor performance of highly purified thiophene oligomers. *Chem. Mater.* **7**, 2235–2237 (1995).
14. Filas, R. W. in *Advances in Electronic Packaging* Vol. 1, 1265–1282 (Am. Soc. Mech. Eng., New York, 1997).
15. Rabay, J. M. in *Digital Integrated Circuits: A Design Perspective* Ch. 6 (Prentice Hall, Upper Saddle River, New Jersey, 1996).
16. Dimitrakopoulos, C. *et al.* Low-voltage organic transistors on plastic comprising high dielectric constant gate insulators. *Science* **283**, 822–824 (1999).
17. Dodabalapur, A., Katz, H. E., Torsi, L. & Haddon, R. C. Organic heterostructure field-effect transistors. *Science* **269**, 1560–1562 (1995).
18. Dodabalapur, A., Baumbach, J., Baldwin, K. & Katz, H. E. Hybrid organic/inorganic complementary circuits. *Appl. Phys. Lett.* **68**, 2246 (1996).
19. Bonse, M. *et al.* in *IEDM Technical Digest* 249–252 (Inst. Elect. Electron. Eng., Piscataway, New Jersey, USA, 1998).

Acknowledgements

We thank B. Batlogg, E. A. Chandross, A. J. Lovinger, J. H. O'Neill, M. Pinto, V. R. Raju, E. Reichmanis, J. Rogers, R. E. Slusher and P. Wiltzius for discussions.

Correspondence and requests for materials should be addressed to A.D. (e-mail: ananth@bell-labs.com).

Fractal stream chemistry and its implications for contaminant transport in catchments

James W. Kirchner*, Xiaohong Feng† & Colin Neal‡

*Department of Geology and Geophysics, University of California, Berkeley, California 94720-4767, USA

† Department of Earth Sciences, Dartmouth College, Hanover, New Hampshire 03755, USA

‡ Institute of Hydrology, McLean Building, Wallingford, Oxon OX10 8BB, UK

The time it takes for rainfall to travel through a catchment and reach the stream is a fundamental hydraulic parameter that controls the retention of soluble contaminants and thus the downstream consequences of pollution episodes^{1,2}. Catchments with short flushing times will deliver brief, intense contaminant pulses to downstream waters, whereas catchments with longer flushing times will deliver less intense but more sustained contaminant fluxes. Here we analyse detailed time series of chloride, a natural tracer, in both rainfall and runoff from headwater catchments at Plynlimon, Wales. We show that, although the chloride concentrations in rainfall have a white noise spectrum, the chloride concentrations in streamflow exhibit fractal $1/f$ scaling over three orders of magnitude. The fractal fluctuations in tracer concentrations indicate that these catchments do not have characteristic flushing times. Instead, their travel times follow an approximate power-law distribution implying that they will retain a long chemical memory of past inputs. Contaminants will initially be flushed rapidly, but then low-level contamination will be delivered to streams for a surprisingly long time.

A catchment is characterized by a distribution of travel times, reflecting the diverse flow paths that rainfall can take to the stream. Quantifying this travel time distribution is essential for predicting the transport of water and solutes, including soluble contaminants. Catchments are spatially complex and subsurface flow is invisible, so one can only infer the movement and mixing of waters from the isotopic and chemical tracers that they carry. Chloride is an effective chemical tracer³, because it is nonreactive under typical catchment

conditions. Our three small (1.0–3.5 km²) study catchments⁴ are close to the Welsh coastline, and sea-salt chloride inputs fluctuate greatly from one storm to the next. By comparing the chloride signatures of rainfall and runoff through time, we can measure how long the catchments retain a chemical memory of the rain that has fallen in them^{5,6}. We measured chloride in rainfall and streamflow at two different temporal resolutions: weekly for 14 years (780 data points), and daily for over three years (1,140 data points). In both data sets, rainfall concentrations are volume-weighted averages over the preceding day or week, whereas streamflow concentrations are instantaneous values at the time of sampling.

The daily hydrograph (Fig. 1a) shows that storm rainfall inputs are usually matched by prompt (and roughly equal) changes in streamflow. By contrast, streamflow chloride response to storm inputs is strongly damped (Fig. 1b), indicating that peak flows consist mostly of pre-storm water released from the catchment, rather than rainfall flowing directly into the stream⁶ (storm response at Plynlimon, as at many forested catchments, is dominated by groundwater flow and interflow, with return flow occurring at the bases of some slopes^{7–9}). This rapid release of ‘old’ water^{10,11} implies that our catchments transmit hydrologic signals from rainfall to runoff much more rapidly and distinctly than chemical signals. This decoupling of timescales occurs because fluctuations in hydraulic head (which drives water to the stream) can propagate much faster than the water itself¹².

The timescales of catchment hydrologic and chemical response can be clarified using spectral methods^{13,14}, which decompose the rainfall and streamflow signals into their component wavelengths. By comparing the spectral power of the input (rainfall) and output (streamflow) at each wavelength, we can determine how strongly the catchment attenuates hydrologic and chemical signals on each timescale. The spectra of water fluxes in rainfall and streamflow lie nearly on top of one another (Fig. 2a), indicating that all three catchments transmit hydrologic signals with little damping, on all but the shortest timescales.

By contrast, the spectra of the streamflow chloride concentrations (Fig. 2b) show strong attenuation on all wavelengths shorter than 5–10 years, with progressively greater attenuation at shorter wavelengths. Except for a strong annual peak (and its subharmonics) reflecting seasonal fluctuations, the rainfall chloride spectrum scales roughly as white noise. The streamflow chloride spectra, by contrast, show fractal power-law scaling that resembles $1/f$ noise (Fig.

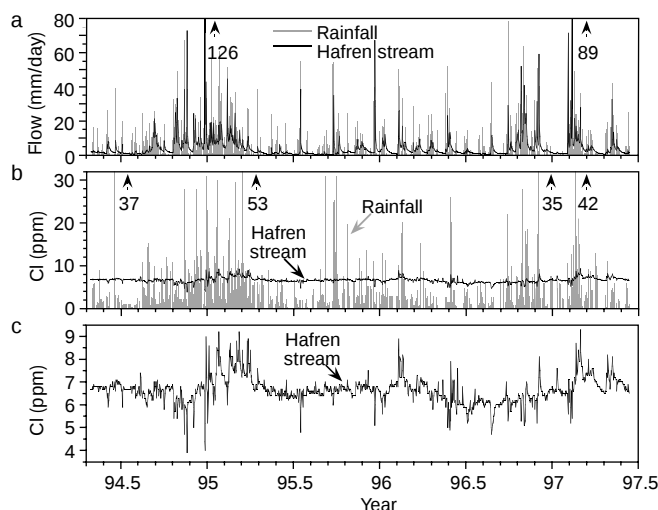


Figure 1 Time series of daily water fluxes and chloride concentrations. **a, b**, At the Hafren catchment, Plynlimon, Wales, hydrologic response to rainfall inputs (**a**) is prompt, with little attenuation, but chemical response (**b**) is strongly damped. Peaks beyond plot axes

are labelled individually. **c**, Enlarged view of the stream chloride time series shows that it exhibits both short-term transience and long-term persistence. For details of measurement methods see ref. 3.

2b), with spectral power inversely proportional to frequency; this indicates that catchments can act as fractal filters, converting white noise inputs into $1/f$ noise outputs. The fractal scaling in the streamflow chloride spectra is consistent between the daily and weekly data sets, consistent among five sampling sites in our three study catchments (see Supplementary Information), and consistent across timescales spanning three orders of magnitude—a wider range of scales than is usually used to characterize fractal behaviour¹⁵.

We have also tested the generality of these results using long (8–30 year) time series of weekly chloride measurements in nine Scandinavian and North American streams; eight of the nine exhibit fractal power-law scaling, between $1/f^{0.7}$ and $1/f^{1.2}$, over 1.5–2.5 orders of magnitude. Whereas the Plynlimon catchments are underlain by deeply fractured shale, the Scandinavian and North American catchments mostly lie on glaciated granitic and metamorphic rocks and thus are much less permeable at depth. The prevalence of fractal scaling means that it does not require unusual catchment characteristics.

The damping of tracer fluctuations from rainfall to streamflow can be used to infer a catchment's travel time distribution^{16,17}. Each day's rainfall influences stream chemistry in future days, by an amount that is determined by the travel time distribution. Equivalently, the present concentration in the stream reflects the rainfall concentrations throughout the past, weighted by their fractional contribution to the present runoff. Mathematically this means that the stream concentration $c_s(t)$ at any time t is the convolution of the travel time distribution $h(\tau)$ and the rainfall concentration $c_R(t - \tau)$

throughout the past, where τ is the lag time between rainfall and runoff:

$$c_s(t) = \int_0^\infty h(\tau)c_R(t - \tau)d\tau \quad (1)$$

Because the flow rate varies through time, equation (1) is strictly valid when t and τ are expressed in terms of the cumulative flow through the catchment, rather than calendar time^{18,19}, but the mathematics are the same in either case¹⁹. We performed the following analysis both ways and obtained functionally equivalent results, so for simplicity we use the time-based formalism below.

From equation (1) one can see that if a catchment's travel time distribution is broad compared to the wavelength of a chemical signal in the rainfall, signals travelling by different flow paths will tend to average each other out when they reach the stream. Thus the average concentration will be maintained, but fluctuations will be

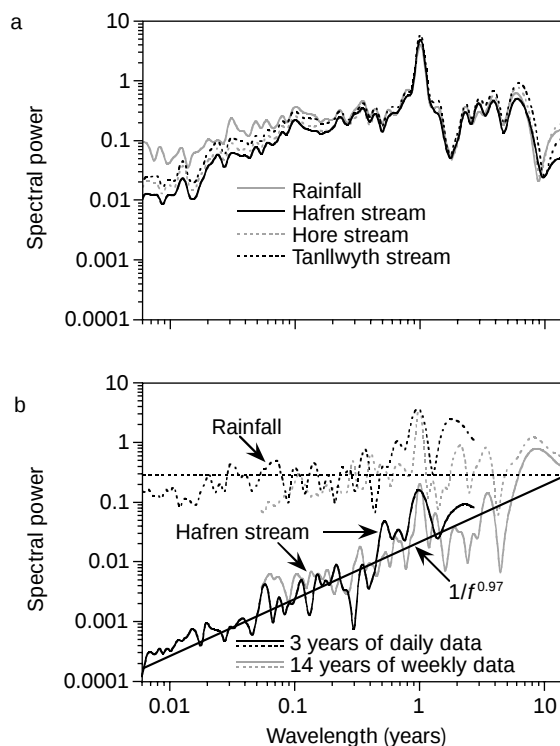


Fig. 2 Power spectra of water fluxes and chloride concentrations. Water flux (rainfall and streamflow) spectra (**a**) indicate little damping, except at the shortest timescales, whereas chloride concentration spectra (**b**) show strong attenuation, except at the longest timescales. Chloride spectra of rainfall (dotted lines) resemble white noise; those of streamflow (solid lines) resemble $1/f$ noise, with spectral power increasing proportionally to wavelength across the entire range of scales. For ease of interpretation, spectra are shown as functions of wavelength (reciprocal of frequency) rather than frequency. **a**, Data are daily for 14 years; **b**, data are daily for 3 years, or weekly for 14 years (see figure). Hafren, Hore and Tanllwyth are adjacent catchments at Plynlimon, Wales.

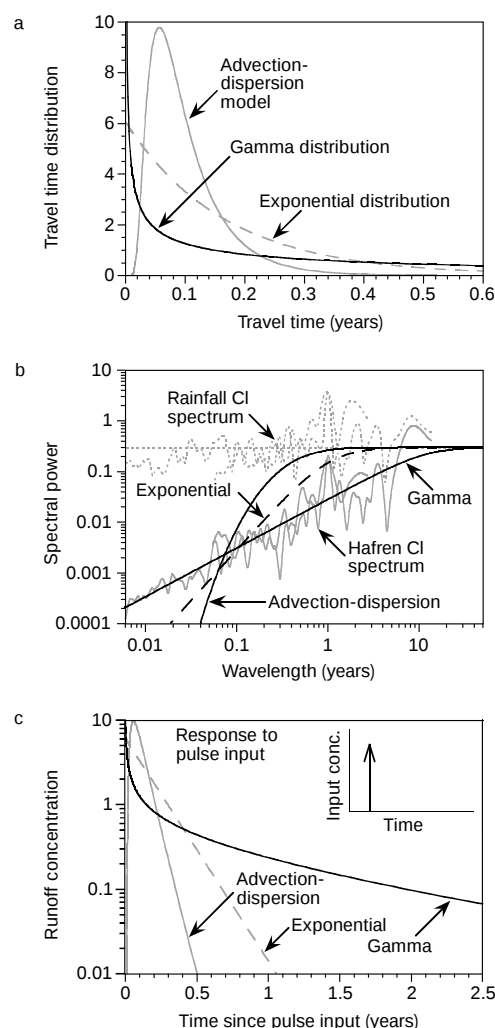


Figure 3 Travel time distributions and their implications for contaminant transport. **a**, Power-law (gamma) distribution compared to two conventional models for catchment travel times (advection–dispersion and exponential). **b**, Power spectra of the three distributions, superimposed on the rainfall and Hafren chloride concentration spectra. Modelled power spectra assume that the rainfall concentration spectrum is white noise (straight dotted line). **c**, Response of streamflow concentrations to a delta-function pulse input of contaminants. Because the gamma distribution has a much longer tail than the conventional models, it sustains substantial contaminant concentrations for much longer time spans. The logarithmic concentration scale emphasizes the persistence of low-level contamination. Inset depicts delta-function contaminant input.

attenuated. The shorter the wavelength of the chemical fluctuation in rainfall compared to the catchment's travel time distribution, the stronger this attenuation by averaging becomes. Conversely, chemical variations on timescales that are long compared to the travel time distribution will be transmitted through the catchment without significant attenuation. By the convolution theorem, equation (1) implies that²⁰:

$$C_S(f) = H(f)C_R(f) \quad \text{and} \quad |C_S(f)|^2 = |H(f)|^2|C_R(f)|^2 \quad (2)$$

Here f is frequency; $C_S(f)$, $H(f)$ and $C_R(f)$ are the Fourier transforms of $c_S(t)$, $h(\tau)$ and $c_R(t - \tau)$; and $|C_S(f)|^2$, $|H(f)|^2$ and $|C_R(f)|^2$ are their power spectra. In our case, because the rainfall concentration spectrum is nearly white noise ($|C_R(f)|^2$ is approximately constant), the spectrum of the travel time distribution is roughly proportional to the runoff concentration spectrum ($|H(f)|^2 \propto |C_S(f)|^2$). From equation (2) it is clear that the power-law scaling observed in Fig. 2b can arise only if the travel time distribution is an approximate power function of the form $h(\tau) \propto \tau^{-m}$ with $m \approx 0.5$. It is equally clear that this approximation must break down for large τ , because otherwise the average travel time would become infinite. (No such difficulty occurs for small τ , since τ^{-m} ($0 < m < 1$) is integrable, though unbounded, at $\tau = 0$.) An approximate power function that is integrable at large τ is the gamma distribution (Fig. 3a):

$$h(\tau) = \frac{\tau^{\alpha-1}}{\beta^\alpha \Gamma(\alpha)} e^{-\tau/\beta} \quad (3)$$

Here β is a scale parameter and $\alpha = 1 - m \approx 0.5$ is a shape parameter. With a fitted value of $\alpha = 0.48 \pm 0.01$, the gamma distribution's power spectrum²¹

$$|H(f)|^2 = (1 + 4\pi^2 f^2 \beta^2)^{-\alpha} \quad (4)$$

agrees well with the observed power spectrum (Fig. 3b), scaling as $f^{-2\alpha} \approx f^{-0.97}$ for $2\pi f\beta \gg 1$. The gamma distribution has previously been used to model catchment travel times²², but with $\alpha \geq 1$ rather than $\alpha \approx 0.5$, for which it has a completely different shape.

Figure 3 contrasts the gamma distribution with two distributions that are often used to model travel times in catchments. If catchments are modelled as well-mixed reservoirs, their travel times follow an exponential distribution^{16,18,22}, which is equivalent to equation (3) with $\alpha = 1$. Thus a well-mixed reservoir also exhibits power-law scaling, but with a non-fractal slope of 2, rather than the observed fractal slope of ~ 1 (Fig. 3b). Alternatively, if solutes are assumed to reach the stream by transport and dispersion along a single flow path, their travel times are described by the advection–dispersion equation^{23–25}, whose power spectrum¹⁶ shows no power-law scaling (Fig. 3b). Both of these alternative models strongly attenuate fluctuations below some characteristic timescale; thus they will either greatly overestimate the attenuation of short-term fluctuations, or greatly underestimate the timescales required for output concentrations to track inputs (and thus the flushing time for the catchment). The stream concentration time series (Fig. 1c) and spectrum (Fig. 2b) indicate that the catchment exhibits both short-term responsiveness to input fluctuations, and long-term memory of past inputs. Only the approximate power-law distribution (equation (3)) has both of these characteristics.

Our finding that tracer fluctuations in streams are fractals, and thus that catchment travel time distributions are approximate power functions, has important implications for contaminant retention by catchments. Conventional methods for predicting cleanup timescales assume that solute travel times follow either an exponential or advection–dispersion distribution²⁶. Compared to those distributions, equation (3) is much steeper at very short lags, but converges toward zero very slowly for long lag times (Fig. 3a). This implies that after pollution episodes, contaminant concentrations will fall rapidly at first, but then decline much more gradually

(Fig. 3c). Expectations based on the initial rapid drop will be thwarted by the persistence of contaminant concentrations over the long term. This persistence may be further amplified by adsorption/desorption reactions between contaminants and mineral surfaces (“retardation”), which do not affect our chloride tracer.

Soil pores and bedrock fractures exhibit fractal scaling^{27,28}, which may be related to the power-law distribution of travel times and the fractal structure of stream chemistry fluctuations. However, catchment travel time distributions are also shaped by many other factors: the spatial distribution of rainfall, the length, tortuosity, and velocity of the flowpaths connecting the stream to each point on the surface, dispersion along and between flowpaths, and diffusive exchange between mobile water in macropores and immobile water in the matrix between them²⁴. The travel time distribution (equation (3)) and power spectra (Fig. 2) presented here provide important observational constraints on the physical processes controlling water flow, storage, and mixing at catchment scale. Understanding these processes, and their implications for water quality, is a critical challenge in the environmental sciences²⁹. □

Methods

Power spectra

Because rainfall chloride concentrations are undefined on dates without rain, we calculated power spectra using the Lomb–Scargle Fourier transform¹⁴, which has the same statistical properties as the conventional Fast Fourier Transform, but does not require evenly spaced data. We omitted Scargle's prefactor of $(N_0/2)^{1/2}$ in order to make spectral power comparable between the daily and weekly data sets, which have different numbers of points.

Travel time distributions

The power spectra of all three travel time distributions were fitted to the stream chloride spectra (Fig. 3b) by least squares. The gamma distribution is equation (3), with parameter values of $\alpha = 0.48 \pm 0.01$ and $\beta = 1.9 \pm 0.1$ years (corresponding to an average travel time of $\alpha\beta = 0.9$ years). The power-law slope of its spectrum is controlled by α , and its left-to-right location is controlled by β . The exponential distribution is equation (3) with α fixed at 1 and a fitted mean travel time of $\alpha\beta = 0.16 \pm 0.01$ years. The advection–dispersion model is equation (11) of Kreft and Zuber²³, with an assumed Peclet number of 5 and a fitted mean travel time of 0.10 ± 0.01 years; other realistic Peclet numbers yield equally poor fits to the data.

Received 24 May; accepted 1 November 1999.

- Langmuir, D. *Aqueous Environmental Geochemistry* (Prentice-Hall, Upper Saddle River, New Jersey, 1997).
- Schnoor, J. L. *Environmental Modeling: Fate and Transport of Pollutants in Water, Air, and Soil* (Wiley, New York, 1996).
- Neal, C. & Rosier, P. T. W. Chemical studies of chloride and stable oxygen isotopes in 2 conifer afforested and moorland sites in the British uplands. *J. Hydrol.* **115**, 269–283 (1990).
- Neal, C. *et al.* The hydrochemistry of the River Severn, Plynlimon. *Hydrol. Earth Syst. Sci.* **1**, 583–617 (1997).
- Robson, A., Neal, C., Hill, S. & Smith, C. J. Linking variations in short-term and medium-term stream chemistry to rainfall inputs – some observations at Plynlimon, mid-Wales. *J. Hydrol.* **144**, 291–310 (1993).
- Neal, C. *et al.* Chloride in precipitation and streamwater for the upland catchment of the River Severn, mid-Wales: some consequences for hydrochemical models. *Hydrol. Process.* **2**, 155–165 (1988).
- Kirkby, C., Newson, M. D. & Gilman, K. *Plynlimon research: the first two decades* (Institute of Hydrology Report 109, Institute of Hydrology, Wallingford, UK, 1991).
- Neal, C., Hill, T. & Reynolds, B. Acid neutralisation capacity measurements in surface and groundwaters in the upper River Severn, Plynlimon: from hydrograph splitting to water flow pathways. *Hydrol. Earth Syst. Sci.* **1**, 687–696 (1997).
- Neal, C. *et al.* The occurrence of groundwater in the lower Palaeozoic rocks of upland Wales. *Hydrol. Earth Syst. Sci.* **1**, 3–18 (1997).
- Buttle, J. M. Isotope hydrograph separations and rapid delivery of pre-event water from drainage basins. *Prog. Phys. Geog.* **18**, 16–41 (1994).
- Kish, M. G., Beven, K. J., Gilman, K. & Darling, W. G. Isotope studies of pipeflow at Plynlimon, Wales, UK. *Hydrol. Process.* **10**, 921–944 (1996).
- Beven, K. On subsurface stormflow: predictions with simple kinematic theory for saturated and unsaturated flows. *Wat. Resour. Res.* **18**, 1627–1633 (1982).
- Bracewell, R. N. *The Fourier Transform and its Applications* 3rd edn (McGraw-Hill, Boston, 2000).
- Scargle, J. D. Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data. *Astrophys. J.* **263**, 835–853 (1982).
- Avnir, D., Biham, O., Lidar, D. & Malcai, O. Is the geometry of nature fractal? *Science* **279**, 39–40 (1998).
- Duffy, C. J. & Gelhar, L. W. A frequency domain approach to water quality modeling in groundwater: theory. *Wat. Resour. Res.* **21**, 1175–1184 (1985).
- Duffy, C. J. & Gelhar, L. W. A frequency domain analysis of groundwater quality fluctuations: interpretation of field data. *Wat. Resour. Res.* **22**, 1115–1128 (1986).

18. Rodhe, A., Nyberg, L. & Bishop, K. Transit times for water in a small till catchment from a step shift in the oxygen-18 content of the water input. *Wat. Resour. Res.* **32**, 3497–3511 (1996).
19. Niemi, A. J. Residence time distributions of variable flow processes. *Int. J. Appl. Radiat. Isotopes* **28**, 855–860 (1977).
20. Gelhar, L. W. *Stochastic Subsurface Hydrology* (Prentice-Hall, Englewood Cliffs, NJ, 1993).
21. Bain, L. in *Encyclopedia of Statistical Sciences* (eds. Kotz, S. & Johnson, N. L.) 292–298 (Wiley, New York, 1983).
22. Turner, J. V. & Barnes, C. J. in *Isotope Tracers in Catchment Hydrology* (eds Kendall, C. & McDonnell, J. J.) 723–760 (Elsevier, Amsterdam, 1998).
23. Kreft, A. & Zuber, A. On the physical meaning of the dispersion equation and its solutions for different initial and boundary conditions. *Chem. Eng. Sci.* **33**, 1471–1480 (1978).
24. Sposito, G., White, R. E., Darrah, P. R. & Jury, W. A. A transfer function model of solute transport through soil 3. The convection-dispersion equation. *Wat. Resour. Res.* **22**, 255–262 (1986).
25. Stewart, M. K. & McDonnell, J. J. Modeling base flow soil water residence times from deuterium concentrations. *Wat. Resour. Res.* **26**, 3005–3019 (1991).
26. Kavanaugh, M. C. et al. *Alternatives for Ground Water Cleanup* (National Academy Press, Washington, DC, 1994).
27. Turcotte, D. L. *Fractals and Chaos in Geology and Geophysics* (Cambridge University Press, 1992).
28. Korvin, G. *Fractal Models in the Earth Sciences* (Elsevier Science, Amsterdam, 1992).
29. Neal, C. A view of water quality from the Plymilton watershed. *Hydrol. Earth Syst. Sci.* **1**, 743–754 (1997).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

Our collaboration was supported by a National Science Foundation grant to J.W.K. Sample collection and analysis were supported by the Natural Environment Research Council, the Environment Agency of England and Wales, and the Forestry Commission. We thank the Plymilton field staff for sample collection, M. Neal for sample analysis, C. Stark for comments on the manuscript, and D. Brillinger for helpful advice.

Correspondence and requests for materials should be addressed to J.W.K. (e-mail: kirchner@seismo.berkeley.edu).

Rapid diffusive infiltration of sodium into partially molten peridotite

Craig C. Lundstrom

Department of Geological Sciences, Brown University, Providence, Rhode Island 02912, USA

Present address: Department of Geology, University of Illinois, Urbana, Illinois 61801, USA.

Recent seismological, geochemical and experimental observations suggest that, as mantle peridotite melts, the resulting basaltic liquid forms an interconnected network, culminating in the rapid ascent of the basalt relative to the surrounding solid matrix^{1–3}. Mantle melting is therefore a polybaric process, with melts produced over a range of pressures having differing chemical characteristics^{4–6}. Modelling and peridotite-melting experiments designed to simulate polybaric mantle melting generally assume that there is no interaction between melts generated at greater pressures and the overlying solid mantle at lower pressures^{5,7}. Beneath mid-ocean ridges, melts derived from greater depth are probably channelized during ascent, so preventing direct re-equilibration with shallow peridotite⁸, as required by geochemical observations^{6,9}. I show here, however, that sodium in ascending melts will quickly diffuse into the melt formed within nearby peridotite at lower pressures. This process fundamentally changes the manner by which the peridotite melts, and can account for both the creation of silica-rich glass inclusions in mantle xenoliths and the anomalous melting modes recorded by abyssal peridotites. Increased melting of lithosphere and upwelling asthenosphere could result from this process without the need to invoke higher mantle temperatures.

On the basis of experimental results, early formed high-pressure

melts derived from the base of a melting column will be lower in SiO₂ and higher in Na₂O than larger-degree melts in equilibrium with mantle peridotite at shallow depths^{10,11}. To simulate the interaction between a melt originally generated at greater pressures and peridotite at lower pressure, a two-step experiment was performed in a piston cylinder apparatus at 1,300 °C and 0.9 GPa using salt–Pyrex assemblies (10% friction correction). The initial synthesis runs (30 hours) produced (1) a peridotite coexisting with ~10% tholeiitic glass (starting material was spinel therzolute KLB-1¹²) and (2) a basaltic glass coexisting with ~10% olivine (starting material was EL-10 from the Canary Islands, collected by K. Hoernle), both enclosed in graphite inner capsules sealed inside Pt-lined Mo outer capsules. After synthesis, a portion of each of these materials was removed for analysis. The capsules were then each polished flat on one end and juxtaposed in the piston cylinder for 10 minutes under the same conditions of temperature and pressure. The peridotite–basaltic interface after reaction remained well defined and fixed in its original location; however, peridotite close to the interface contained more melt than did peridotite far from the interface (Fig. 1).

Compositions of melt pools within the peridotite vary as a

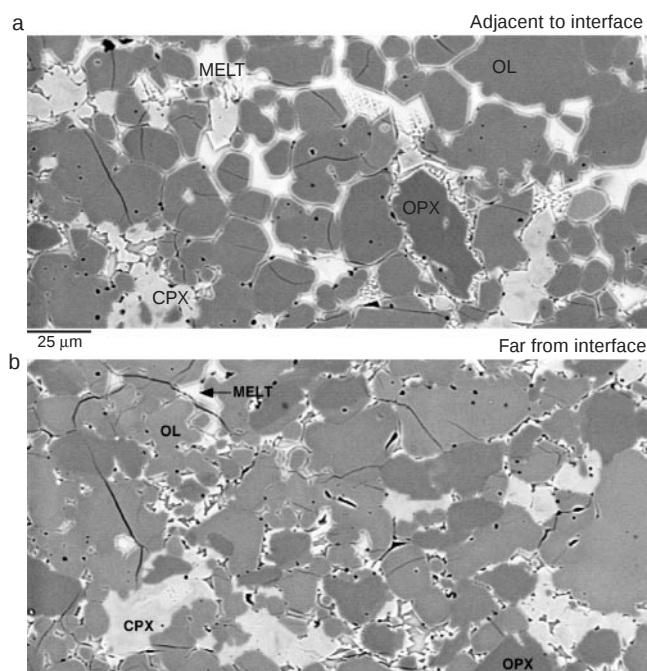


Figure 1 Back-scattered electron (BSE) images of peridotite. **a**, Closest to the basaltic–peridotite interface; **b**, farthest from the interface. Note the clear difference in melt and mineral modes between the two regions of the experiment. Representative minerals are labelled (orthopyroxene (OPX), dark grey; olivine (OL), medium grey; clinopyroxene (CPX) and MELT, both light grey) with lighter jagged edges on pyroxenes and lighter rims on olivine resulting from quench crystallization. In order to quantify phase proportions in the experiment (Fig. 3), four non-overlapping, digital BSE images (250 μm²) and Mg X-ray maps were simultaneously collected throughout the peridotite portion of the sample in each of four rows that varied as a function of distance from the original peridotite–basaltic interface (see Fig. 2 for locations of rows). Each image was processed using NIH Image software with mineral modes calculated by summing areas of particular grey scale. With the aid of Mg X-ray maps, melt, orthopyroxene and clinopyroxene were distinguished from olivine and the mode of each was counted after selectively darkening phases using imaging software. For each row, an average mode and 1 standard deviation (s.d.) were calculated from the four images. Quench crystals could be readily distinguished from the primary minerals; the melt mode reflects the area covered by both melt and quench crystals. Cracks and surface imperfections (black) are not included in the modal determination.

function of distance from the interface (Fig. 2). The compositional profiles of most elements are similar to those expected for binary diffusion between the tholeiite in equilibrium with the peridotite and the basanite. In contrast, Na_2O rapidly diffuses into the peridotite, abruptly increasing in concentration in melt pools near the interface and remaining well above its starting-material concentration for 1 mm into the peridotite. Despite greater initial Na_2O contents in the basanite, concentrations evolve into an 'uphill' profile in 10 minutes, with the Na_2O content of the basanite dropping from 4 to 3.4 wt %. K_2O also has a longer diffusion profile, indicating fast diffusion relative to other elements. Although SiO_2 does not diffuse quickly, its concentration profile mimics that of Na_2O .

Intergranular melts in peridotite melting experiments, including this one, are modified by crystallization during quenching. However, quench modification cannot account for the elemental variations observed. First, the peaks in SiO_2 and Na_2O occur in the area having the largest amount of melt; this area should be

the least affected by quench modification. Second, the effect of quench-modification shown by the original starting material (peridotite+melt) is small relative to the changes in element profiles after infiltration (Fig. 2). Third, time-series experiments that reacted a 2:1 intimate mixture of KLB-1 and EL-10 at identical pressures and temperatures result in quench-modified melts containing only 51 wt% SiO_2 .

Mineral modes (volume percentages) and mineral compositions within the peridotite vary as a function of distance from the interface (Fig. 3). Approaching the interface, the orthopyroxene mode decreases while the melt mode increases. The increased melt near the interface does not reflect bulk infiltration of the basanite followed by reaction with the peridotite for the following reasons. The low TiO_2 and P_2O_5 contents of melts near the interface restrict the amount of basanite that could have infiltrated to less than 25%, whereas the melt mode has in fact doubled. Second, the low SiO_2 content of the melt in the intimate-mixture experiment shows that basanite–pyroxene reaction does not produce high- SiO_2 melts. Last, the buoyancy or capillary forces present in the experiment will not drive the basanite into the peridotite. Instead, greater amounts of melt reflect increased peridotite melting due to the diffusive infiltration of Na_2O increasing the bulk content of Na_2O to ~0.8 wt % within the infiltrated portion of peridotite.

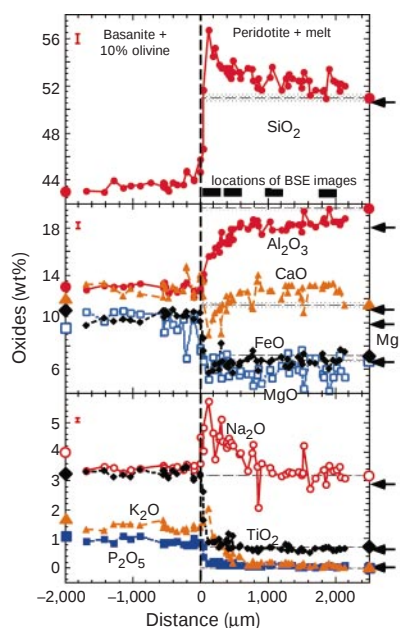


Figure 2 Elemental profiles as a function of distance from the original basanite–peridotite interface ($x=0$). Data are shown for the basanite and within melt pools in the peridotite. Analysed melt pools within the peridotite range in size from 10 to 40 μm . Dashed horizontal lines over grey indicate the average ± 1 s.d. of the melt-pool composition within the starting peridotite–melt aggregate, with symbols on the right indicating values for respective elements. Arrows on the right point to non-quench-modified melt composition of the peridotite+melt starting material, determined by using a small layer of vitreous carbon during the synthesis run. Note the relatively small change in element concentrations due to quench crystallization. The change in MgO between the equilibrium melt and the quench-modified melt corresponds to ~15% crystallization of Fe_{78} olivine. Symbols on the left axis show the concentrations of elements in the basanite glass starting material. Small boxes show the width and location of the BSE images. Na_2O produces a pronounced peak within the peridotite reflecting the diffusive infiltration of Na_2O , and is markedly depleted in the basanite relative to its starting value (by 0.6%). Mass balance calculated using the modal abundance of melt shows all Na_2O to be accounted for by this distribution. The corresponding peak in SiO_2 does not result from diffusion but rather reflects the preferential melting of orthopyroxene due to Na_2O infiltration. Electron microprobe (Brown University) analysis conditions were as follows: 6–8 μm diameter beam, 10 nA current and 15 kV accelerating voltage were used for analysing glasses, while minerals were analysed using a 15 nA current. Natural mineral standards were used, and melt compositions were normalized to values for basaltic standard VG-2 collected before and after analysis. Representative 1 s.d. errors are given in the upper left of each panel.

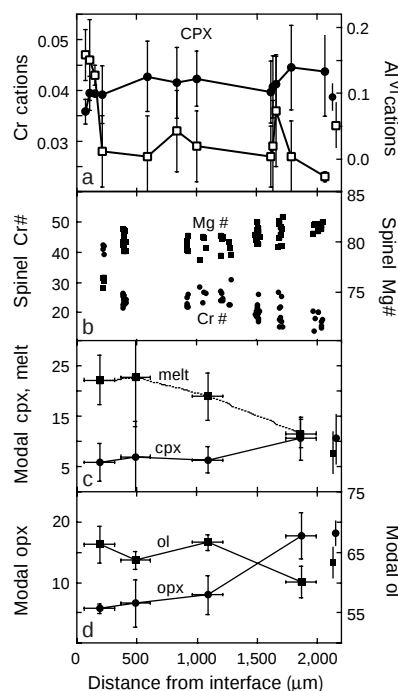


Figure 3 Variation in clinopyroxene and spinel compositions and mineral modes with distance from the original basanite–peridotite interface. **a**, Clinopyroxene Cr content (open squares) increases while octahedral Al content (filled circles) decreases approaching the interface. Each symbol represents the average $\pm$ s.d. of 3–15 analyses in a given area. Symbols on the far right indicate starting material composition. **b**, Spinel Cr# ($\text{Cr}/(\text{Cr}+\text{Al})$; filled circles) increases and Mg# ($\text{Mg}/(\text{Mg}+\text{Fe}^{2+})$; filled squares) decreases close to the interface, consistent with changes in spinel composition being coupled to incongruent melting of orthopyroxene¹³. **c,d**, The orthopyroxene mode decreases and the melt mode increases as the interface is approached, reflecting the infiltration of Na_2O causing preferential consumption of orthopyroxene. Although the peridotite is modally heterogeneous at the scale of individual images, the most important observation, that orthopyroxene is preferentially consumed close to the interface, is clearly resolved in these results. Errors on modes are 1 s.d. of the four image analyses for each row, and horizontal error bars reflect width of BSE image. Modes with error bars in the starting material after the synthesis step are shown at far right.

Although zoned, average spinel compositions change from high-Mg# ($\text{Mg}/(\text{Mg}+\text{Fe}^{2+})$), aluminous spinels (typical of the starting material) to lower-Mg# and higher-Cr# ($\text{Cr}/(\text{Cr}+\text{Al})$) spinels approaching the interface (Fig. 3). This inverse Cr#–Mg# relationship could reflect coupling between spinel reactions and incongruent melting of orthopyroxene¹³. Spinel changes from an anhedral morphology (typical of the starting material) to a small, euhedral spinel ($\sim 8\text{-}\mu\text{m}$ diameter) near the interface. Clinopyroxenes, also zoned, become more refractory (more Cr-rich, less aluminous) near the interface. In contrast, olivine and orthopyroxene are relatively homogenous, and show no significant compositional difference between crystals near to and far from the interface.

To gain insight into the diffusive infiltration of sodium into peridotite+melt, I performed a melt–melt diffusion experiment using presynthesized polished capsules of EL-10 and a tholeiitic melt (A2384-1; a high-MgO, primitive mid-ocean-ridge basalt (MORB) glass from the Siqueiros transform) for 5 minutes at 0.9

GPa and $1,450^\circ\text{C}$ (Fig. 4). The resulting ‘humped’ Na_2O profile indicates a strong coupling of Na_2O diffusion to gradients in SiO_2 . This is in accord with experimental studies showing sodium to diffuse rapidly from melts of lower silica content into melts of higher silica content, including ‘uphill diffusion’ of Na_2O , until equilibrium two-melt partitioning is established¹⁴. Although the two-melt partition coefficient between basanite and tholeiite is unknown, Na_2O clearly prefers the more silicic tholeiitic melt, which is consistent with results from basalt–granite and basalt–rhyolite interaction experiments^{14–16}.

The diffusive addition of sodium to the peridotite fundamentally alters how the peridotite melts. Increased alkali contents lower the activity coefficient of silica, such that melts coexisting with olivine and orthopyroxene will have higher SiO_2 contents if Na_2O contents increase^{17–19}. Small-degree melting experiments on peridotite illustrate this effect, producing elevated Na_2O contents and corresponding elevated SiO_2 contents^{5, 20–22}. Thus, the increase in silica content of the melt and the decrease in orthopyroxene mode within the diffusively infiltrated peridotite are consistent with previously observed effects of higher alkali contents on melt chemistry. Combining the SiO_2 dependence of Na_2O diffusion with the enhanced melting of orthopyroxene due to increased sodium concentration results in a positive feedback that should accelerate the diffusive infiltration process. Na_2O infiltration increases the melt SiO_2 content, which in turn increases the rate of Na_2O infiltration. This feedback suggests that reactive diffusive infiltration may be significantly faster than that implied by the melt–melt Na_2O diffusion measurement.

The infiltration experiment documents a process that may exercise profound control on polybaric melting regimes in the mantle. Given the SiO_2 contents of peridotite partial melts as a function of pressure, sodium will diffusively migrate from melt ascending in conduits to melt within adjacent peridotite. Sodium infiltration fundamentally alters how the peridotite melts by increasing the contribution of orthopyroxene to the melt, providing an explanation for several enigmatic observations of mantle melting.

Glasses high in silica, alumina and alkalis are commonly found in peridotite xenoliths entrained in alkali-rich basalts from ocean island and continental volcanic settings worldwide^{23,24}. The origin of these glasses has been controversial, with many suggested explanations^{25,26}. Experiments show that some xenolith glass compositions are in equilibrium with spinel lherzolite near 1 GPa (ref. 26). However, the high Na_2O and K_2O contents of such glasses are often difficult to reconcile with the glasses’ being small-degree melts of the host xenolith, particularly because the highest- SiO_2 glasses are often found in refractory harzburgites (see, for example, ref. 24). Diffusive infiltration of alkalis from alkali-rich host basalts into xenoliths provides a possible explanation for the origin of high-silica glasses. Although the present infiltration experiment results in SiO_2 and Al_2O_3 contents of the melt that are less than those often observed, the experimental temperature is considerably higher than indicated by geothermometry of xenoliths ($\sim 1,150\text{--}1,200^\circ\text{C}$) or experimentally determined multiple saturation of xenolith glasses²⁶. Beyond explaining high-silica glasses in xenoliths, diffusive infiltration of alkalis into peridotite may imply that long-term fluxing of alkali-rich, silica-undersaturated magmas through mantle lithosphere could induce significant lithospheric melting by a process that is chemically, not thermally, activated.

The mineral modes of abyssal peridotites define coherent melting trends, leading to the view that abyssal peridotites are the solid residues that remain after extraction of MORB²⁷. However, the melting trend recorded by abyssal peridotite modes differs from experimentally based predictions; orthopyroxene decreases more than predicted²⁸. Although simplistic, a calculation using feasible parameters indicates that diffusive infiltration of Na_2O from melt

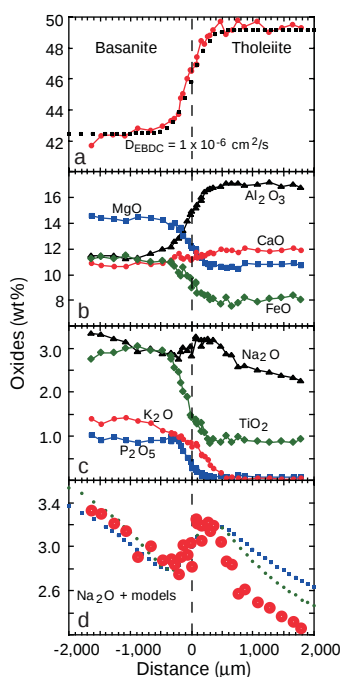


Figure 4 Elemental profiles of a two-melt diffusion couple run for 5 min at $1,450^\circ\text{C}$ and 0.9 GPa. **a–c**, Most elements including SiO_2 produce simple binary diffusion profiles consistent with effective binary diffusion coefficients of $\sim 10^{-6}\text{ cm}^2\text{ s}^{-1}$; a representative model is shown for SiO_2 . However, alkalis deviate from this behaviour, with K_2O showing a diffusion profile consistent with $D \sim 5 \times 10^{-6}\text{ cm}^2\text{ s}^{-1}$ and Na_2O producing a humped profile that is not consistent with simple binary diffusion. **d**, The shape of the Na_2O profile can be modelled by describing Na_2O diffusion as reflecting gradients in both Na_2O and SiO_2 using the following (Y. Liang, personal communication):

$$\frac{\partial C_{\text{Na}}}{\partial t} = D_{\text{Na,Na}} \frac{\partial^2 C_{\text{Na}}}{\partial X^2} + D_{\text{Na,Si}} \frac{\partial^2 C_{\text{Si}}}{\partial X^2}$$

where C is concentration, X is distance from interface and t is time. The Na_2O profile can be modelled using a 2×2 diffusion matrix based on the binary diffusion behaviour of SiO_2 . The modelled profile of Na_2O varies with the value of the diffusion coefficient of Na_2O due to gradients in Na_2O ($D_{\text{Na,Na}}$) and the diffusion coefficient of Na_2O due to gradients in SiO_2 ($D_{\text{Na,Si}}$). Squares represent model with $D_{\text{Na,Na}} = 6 \times 10^{-6}\text{ cm}^2\text{ s}^{-1}$ and $D_{\text{Na,Si}} = -8 \times 10^{-6}\text{ cm}^2\text{ s}^{-1}$. Circles represent model with $D_{\text{Na,Na}} = 3 \times 10^{-6}\text{ cm}^2\text{ s}^{-1}$ and $D_{\text{Na,Si}} = -5 \times 10^{-6}\text{ cm}^2\text{ s}^{-1}$. Both models assume $D_{\text{Si,Si}} = 1 \times 10^{-6}\text{ cm}^2\text{ s}^{-1}$ and $D_{\text{Si,Na}} = 1 \times 10^{-8}\text{ cm}^2\text{ s}^{-1}$. Therefore, the steep gradient in melt SiO_2 content at the interface of both the basanite–tholeiite and the basanite–peridotite experiments drives the rapid diffusion of Na_2O into the higher- SiO_2 melt.

conduits into adjacent peridotite could affect the entire melting region beneath a mid-ocean ridge. If peridotite adjacent to a melt conduit were exposed to alkali diffusive infiltration for ~ 7 Myr (assuming a solid upwelling rate of centimetres per year), the diffusive lengthscale—given by $l = 2(D\phi t)^{0.5}$, where D is diffusion coefficient and ϕ is melt porosity—would be ~ 230 m: this assumes that $\phi = 1\%$ and that D for Na (D_{Na}) is $6 \times 10^{-5} \text{ cm}^2 \text{ s}^{-1}$. Thus, this process could affect melting of all peridotite located between melt conduits spaced less than 0.5 km apart.

Abyssal peridotites near to hotspots have the largest trace-element depletions and the highest olivine/orthopyroxene ratios of all peridotites, this is commonly ascribed to higher degrees of melting and higher mantle temperatures⁶. An alternative explanation implied by the present results is that alkali-rich melts from the hotspot source have fluxed through shallow peridotite. Evidence for the diffusive infiltration process comes from two observations concerning Na_2O . First, although most incompatible-element contents in abyssal peridotites require a fractional melting process⁶, Na_2O contents are higher than expected from such a process²⁹. Second, Na_2O shows much less variability in MORB suites than elements with similar bulk partition coefficients³⁰, which is consistent with the diffusive infiltration process acting to homogenize differences in Na_2O .

The experiments I report here show that diffusive infiltration of sodium will rapidly occur whenever alkali-rich, silica-poor magmas contact peridotite at shallow depth and magmatic temperatures. This process can strongly affect the melting relations of peridotite: it could provide an explanation for the genesis of high-silica melts in spinel lherzolite xenoliths, and could also explain the preferential melting of orthopyroxene in the shallow mantle beneath mid-ocean ridges. □

Received 14 April; accepted 9 December 1999.

- Toomey, D. R., Wilcock, W. S. D., Solomon, S. C., Hammond, W. C. & Orcutt, J. A. Mantle seismic structure beneath the MELT region of the East Pacific Rise from P and S wave tomography. *Science* **280**, 1224–1227 (1998).
- McKenzie, D. The generation and compaction of partially molten rock. *J. Petrol.* **25**, 713–765 (1984).
- Waff, H. S. & Bulau, J. R. Equilibrium fluid distribution in an ultramafic partial melt under hydrostatic stress conditions. *J. Geophys. Res.* **84**, 6109–6114 (1979).
- Klein, E. & Langmuir, C. H. Global correlations of ocean ridge basalt chemistry with axial depth and crustal thickness. *J. Geophys. Res.* **92**, 8089–8115 (1987).
- Kinzler, R. J. & Grove, T. Primary magmas of mid-ocean ridge basalts; 2, Applications. *J. Geophys. Res.* **97**, 6907–6926 (1992).
- Johnson, K. T. M., Dick, H. J. B. & Shimizu, N. Melting in the oceanic upper mantle: an ion microprobe study of diopsides in abyssal peridotites. *J. Geophys. Res.* **95**, 2661–2678 (1990).
- Hirose, K. & Kushiro, I. The effect of melt segregation on polybaric mantle melting: Estimation from the incremental melting experiments. *Phys. Earth Planet. Inter.* **107**, 111–118 (1998).
- Kelemen, P. B., Shimizu, N. & Salters, V. J. M. Extraction of mid-ocean-ridge basalt from the upwelling mantle by focused flow of melt in dunite channels. *Nature* **375**, 747–753 (1995).
- Kelemen, P. B., Hirth, G., Shimizu, N., Spiegelman, M. & Dick, H. J. B. Review of melt migration processes in the adiabatically upwelling mantle beneath oceanic spreading ridges. *Phil. Trans. R. Soc. Lond.* **355**, 283–318 (1997).
- Stolper, E. M. A phase diagram for mid-ocean ridge basalts: Preliminary results and implications for petrogenesis. *Contrib. Mineral. Petrol.* **74**, 13–27 (1980).
- Hirose, K. & Kushiro, I. Partial melting of dry peridotites at high pressures; determination of compositions of melts segregated from peridotite using aggregates of diamond. *Earth Planet. Sci. Lett.* **114**, 477–489 (1993).
- Takahashi, E. Melting of a dry peridotite KLB-1 up to 14 GPa: implication on the origin of peridotitic upper mantle. *J. Geophys. Res.* **91**, 9367–9382 (1986).
- Lundstrom, C. C., Gill, J. & Williams, Q. A geochemically consistent hypothesis for MORB generation. *Chem. Geol.* **162**, 105–126 (2000).
- Watson, E. B. Basalt contamination by continental crust: some experiments and models. *Contrib. Mineral. Petrol.* **80**, 73–87 (1982).
- Watson, E. B. & Jurewicz, S. R. Behavior of alkalis during diffusive interaction of granitic xenoliths with basaltic magma. *J. Geol.* **92**, 121–131 (1984).
- Leshner, C. E. Kinetics of Sr and Nd exchange in silicate liquids—theory, experiments, and applications to uphill diffusion, isotopic equilibration, and irreversible mixing of magmas. *J. Geophys. Res.* **99**, 9585–9604 (1994).
- Kushiro, I. On the nature of silicate melt and its significance in magma genesis: regularities in the shift of the liquidus boundaries involving olivine, pyroxene and silica minerals. *Am. J. Sci.* **275**, 411–431 (1975).
- Ryerson, F. J. Oxide solution mechanisms in silicate melts—systematic variations in the activity coefficient of silica. *Geochim. Cosmochim. Acta* **49**, 637–649 (1985).
- Hirschmann, M. M., Baker, M. B. & Stolper, E. M. The effect of alkalis on the silica content of mantle-derived melts. *Geochim. Cosmochim. Acta* **62**, 883–902 (1998).

- Walter, M. J. & Presnall, D. C. Melting behavior of simplified lherzolite in the system $\text{CaO-MgO-Al}_2\text{O}_3\text{-SiO}_2\text{-Na}_2\text{O}$ from 7 to 35 kbar. *J. Petrol.* **35**, 329–359 (1994).
- Baker, M. B., Hirschmann, M. M., Ghiorso, M. S. & Stolper, E. M. Compositions of near-solidus peridotite melts from experiments and thermodynamic calculation. *Nature* **375**, 308–311 (1995).
- Robinson, J. A. C., Wood, B. J., & Blundy, J. D. The beginning of melting of fertile and depleted peridotite at 1.5 GPa. *Earth Planet. Sci. Lett.* **155**, 97–111 (1998).
- Frey, F. A. & Green, D. H. The mineralogy, geochemistry, and origin of lherzolite inclusions in Victorian basanites. *Geochim. Cosmochim. Acta* **38**, 1023–1059 (1974).
- Vannucci, R., Botazzi, P., Wulf-Pedersen, E. & Neumann, E. R. Partitioning of REE, Y, Sr, Zr and Ti between clinopyroxene and silicate melts in the mantle under La Palma (Canary Islands): implications for the nature of the metasomatic agents. *Earth Planet. Sci. Lett.* **158**, 39–51 (1998).
- Shaw, C. S. J., Thibault, Y., Edgar, A. D. & Lloyd, F. E. Mechanisms of orthopyroxene dissolution in silica-undersaturated melts at 1 atmosphere and implications for the origin of silica-rich glass in mantle xenoliths. *Contrib. Mineral. Petrol.* **132**, 354–370 (1998).
- Draper, D. S. & Green, T. H. P – T phase relations of silicic, alkaline, aluminous mantle-xenolith glasses under anhydrous and C–O–H fluid-saturated conditions. *J. Petrol.* **38**, 1187–1224 (1997).
- Dick, H. J. B., Fisher, R. L. & Bryan, W. B. Mineralogic variability of the uppermost mantle along mid-ocean ridges. *Earth Planet. Sci. Lett.* **69**, 88–106 (1984).
- Niu, Y., Langmuir, C. H. & Kinzler, R. J. The origin of abyssal peridotites: a new perspective. *Earth Planet. Sci. Lett.* **152**, 251–265 (1997).
- Elthon, D. Chemical trends in abyssal peridotites; refertilization of depleted suboceanic mantle. *J. Geophys. Res.* **97**, 9015–9025 (1992).
- Langmuir, C. H. & Hanson, G. An evaluation of major element heterogeneity in the mantle sources of basalts. *Phil. Trans. R. Soc. Lond. A* **297**, 383–407 (1980).

Acknowledgements

I thank P. Kelemen and P. Asimow for comments on the manuscript, and P. Hess, D. Forsyth, Y. Liang, Q. Williams and M. Rutherford for suggestions. I also thank M. Rutherford for use of his laboratory, J. Devine and M. Jercinovic for assistance with the electron microprobes, and K. Hoernle, E. Takahashi and M. Perfit for sample donation. This work was supported by an SGER grant from the NSF.

Correspondence and requests for materials should be addressed to the author at the University of Illinois (e-mail: lundstro@uiuc.edu).

Oxygen-isotope evidence for recycled crust in the sources of mid-ocean-ridge basalts

John M. Eiler*, Pierre Schiano*†‡, Nami Kitchen* & Edward M. Stolper*

* Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, California 91125, USA

† Laboratoire Géochimie–Cosmochimie, IPG Paris, 4 Place Jussieu, 75252, Paris cedex 05, France

‡ Unité Mixte de Recherche 6524 ‘Magmas et Volcans’, Université Blaise-Pascal, 5 rue Kessler, 63038 Clermont-Ferrand cedex, France

Mid-ocean-ridge basalts (MORBs) are the most abundant terrestrial magmas and are believed to form by partial melting of a globally extensive reservoir of ultramafic rocks in the upper mantle¹. MORBs vary in their abundances of incompatible elements (that is, those that partition into silicate liquids during partial melting) and in the isotopic ratios of several radiogenic isotope systems^{2–4}. These variations define a spectrum between ‘depleted’ and ‘enriched’ compositions, characterized by respectively low and high abundances of incompatible elements^{5,6}. Compositional variations in the sources of MORBs could reflect recycling of subducted crustal materials into the source reservoir⁷, or any of a number of processes of intramantle differentiation^{8–10}. Variations in $^{18}\text{O}/^{16}\text{O}$ (principally sensitive to the interaction of rocks with the Earth’s hydrosphere) offer a test of these alternatives. Here we show that $^{18}\text{O}/^{16}\text{O}$ ratios of MORBs are correlated with aspects of their incompatible-element chemistry. These correlations are consistent with control of the oxygen-isotope and incompatible-element geochemistry of MORBs by a

component of recycled crust that is variably distributed throughout their upper mantle sources.

Geochemical variations in MORBs occur in several physical settings. These include a global-scale isotopic signature of enrichment in low latitudes (the 'DUPAL' anomaly centred on the Indian ocean¹¹); enrichments spatially associated with ocean-island basalt volcanic centres¹²; enrichments in near-ridge seamounts¹³; scattered anomalies of enriched MORB (EMORB, defined as having chondrite-normalized La/Sm ratios greater than 1) within ridge segments dominated by normal MORB (NMORB, defined as having chondrite-normalized La/Sm ratios less than or equal to 1; ref. 14); and ubiquitous, minor variations in the composition of NMORBs (for example, ref. 15) that may reflect either weak influences of the other types of enrichment or a separate phenomenon.

Several aspects of the compositional heterogeneity of MORBs (including variations in abundance ratios of highly incompatible elements and in radiogenic-isotope ratios) cannot easily be explained as consequences of variations in extent of melting of a homogeneous source. Instead, these aspects suggest variations in the compositions of their mantle sources. The following mechanisms have been proposed as causes of compositional variations in the mantle and may contribute to heterogeneity in MORBs: (1) variable extents of depletion by prior partial melting (that is, enriched sources could be residual to relatively less melting than depleted sources)^{8,13}; (2) infiltration of depleted mantle by mantle-derived silicate melts or hydrous and/or carbonic fluids rich in incompatible elements⁹; (3) mixing of depleted mantle rocks with enriched, metasomatized portions of continental lithospheric mantle detached from the roots of continents¹⁰; (4) mixing between depleted mantle and crustal rocks, sediments, and/or fluids injected into the mantle by subduction (for example, refs 7, 16).

The first three of these hypotheses involve intramantle fractionations controlled by high-temperature partitioning of elements

between minerals and silicate melts or fluids; it is therefore difficult to discriminate between them. The fourth hypothesis, however, requires material from the crust as the agent of enrichment. Crustal rocks and sediments often have distinctive properties resulting from exposure to the hydrosphere, including increases or decreases in $\delta^{18}\text{O}$ (where $\delta^{18}\text{O} = ((^{18}\text{O}/^{16}\text{O})_{\text{sample}}/0.0020052 - 1)1000$) relative to values typical of mantle peridotites; for example, altered sea-floor basalts and sediments are elevated in $\delta^{18}\text{O}$ by ~ 5 and ~ 15 – 25% respectively, relative to average mantle^{17,18}, whereas gabbroic rocks in the lower oceanic crust are often lower in $\delta^{18}\text{O}$ than mantle rocks by up to $\sim 5\%$ (ref. 17).

Previous studies of oxygen-isotope variability of fresh MORBs measured values of $\delta^{18}\text{O}$ between 5.35 and 7.30‰ (ref. 19; see Supplementary Information). Variations were not found to be significantly correlated with established indicators of geochemical variability in mantle sources of MORBs; it has therefore been generally suspected that oxygen-isotope ratios in fresh MORBs only vary to small extents, because of trace alteration, magmatic differentiation in crustal magma chambers, or fractionations during melting¹⁹.

We analysed $\delta^{18}\text{O}$ values and water contents of 28 MORB glasses from the east-Pacific rise, the mid-Atlantic ridge and the Indian Ocean (Table 1); sampling locations and other geochemical characteristics of these samples (including previous measurements of $\delta^{18}\text{O}$ and water content) are referenced in the Supplementary Information. Analytical techniques used in this study are described in the Methods. These samples exhibit ranges in the isotopic compositions of Sr, Nd, Pb, and Os that span most of the range known in MORBs (refs 5, 6, 20). All but two samples are within the compositional range of NMORBs; the two exceptions have chondrite-normalized La/Sm ratios of 1.04 and 1.05; that is, they are on the border between NMORB and EMORB compositions. Samples were collected from segments of the ridge system distant from regions clearly influenced by hot-spot volcanism (such as

Table 1 Oxygen isotope data and water contents of MORB glasses

Sample	$\delta^{18}\text{O}^*$	$\pm 1\sigma$	± 1 standard error	wt% H ₂ O	Mg number†	La/Sm _N ‡
Atlantic Ocean						
Ridelente DR10	5.50	0.04	0.02	0.23	0.67	1.05
CH77 DR05 105a	5.40	0.01	0.00	0.27	0.60	n.d.
Mapco CH98 DR12	5.41	0.02	0.01	0.24	0.58	0.48
Mapco CH98 DR11	5.45	0.02	0.01	0.27	0.63	0.50
Mapco CH98 DR10	5.49	0.05	0.04	0.27	0.61	0.48
EW 9309 25D	5.81	0.04	0.03	0.17	0.64	0.97
EW 9309 1D	5.47	0.01	0.01	0.23	0.59	0.84
East Pacific Rise						
Naudure ND 51	5.47	0.02	0.02	0.20	0.66	0.60
Naudure DR21-4	5.41	0.04	0.03	0.21	0.61	0.57
Searise 1 DR5-102	5.43	0.02	0.02	0.26	0.60	0.56
Searise 1 Dr04	5.61	0.02	0.02	0.30	0.55	0.70
CHEPR 60	5.43	0.06	0.03	0.17	0.62	0.54
CY 82 27-1	5.51	0.03	0.02	0.30	0.62	0.75
Clipperton DR01-01a	5.51	0.09	0.04	0.22	0.64	n.d.
Venture number 32	5.67	0.03	0.02	0.28	0.53	n.d.
Searise 2 DR07	5.49	0.08	0.04	0.39	0.47	0.53
CYP 78 18-62	5.45	0.03	0.02	0.20	0.65	0.69
CYP 78 18-65	5.54	0.03	0.02	0.24	0.61	n.d.
CYP 78 04-06	5.37	0.01	0.00	0.18	0.60	0.62
CYP 78 06-10	5.42	0.03	0.02	0.16	0.65	0.55
Indian Ocean						
MD57 D 13-7	5.41	0.04	0.03	0.47	0.59	0.64
MD 57 D'10-1	5.46	0.03	0.02	0.43	n.d.	0.85
MD 57 D'10-4	5.61	0.01	0.00	0.29	0.63	1.04
MD 57 D8-1	5.57	0.06	0.04	0.22	n.d.	0.59
MD 23 site 4	5.49	0.05	0.04	0.24	0.71	0.68
MD 37 07-04-D1	5.45	0.05	0.04	0.23	0.61	0.75
MD 34 D3	5.47	0.01	0.01	0.36	n.d.	0.37
MD 34 D4	5.64	0.01	0.00	0.36	0.58	0.85

* $\delta^{18}\text{O} = ((^{18}\text{O}/^{16}\text{O})_{\text{sample}}/^{18}\text{O}/^{16}\text{O}_{\text{SMOW}} - 1) \times 1000$, where $^{18}\text{O}/^{16}\text{O}_{\text{SMOW}} = 0.0020052$. SMOW, standard mean ocean water.

† Mg number based on data from ref. 20 and the relationship: $\text{Mg\#} = \text{Mg}/(0.9 \times \text{Fe} + \text{Mg})$ where Mg and Fe are molar abundances.

‡ Chondrite-normalized La/Sm ratio.

n.d., not determined; σ , standard deviation.

Iceland), although evidence has been used to infer an influence of 'plumes' on otherwise normal ridge segments²¹ and our suite might sample such effects.

We measured a range in $\delta^{18}\text{O}$ for fresh MORB glass of 5.37–5.81‰ (average of 5.50‰). This range is significantly smaller than that observed in previous studies of MORBs¹⁹, although the minimum $\delta^{18}\text{O}$ observed here is indistinguishable from that observed in previous studies. The differences in range, average and maximum $\delta^{18}\text{O}$ between this and previous studies could reflect differences in analytical precision (see Table 1 and Methods), subtle but systematic differences in the extent of sample alteration, or heterogeneity in fresh MORBs not sampled by our suite.

Values of $\delta^{18}\text{O}$ observed in this study are correlated with absolute and relative abundances of incompatible elements (Fig. 1), a feature not present in previous data sets. The best defined of these correlations are between $\delta^{18}\text{O}$ and the concentration of K_2O and the $\text{K}_2\text{O}/\text{TiO}_2$, K/Sr , $\text{K}_2\text{O}/\text{H}_2\text{O}$, and La/Ti ratios; significant but weaker correlations are observed with the La/Sm , K/U and Ba/Th ratios. Relationships between $\delta^{18}\text{O}$ and radiogenic-isotope ratios will not be considered in detail here (see ref. 20 for radiogenic-isotope data); briefly, relatively high $\delta^{18}\text{O}$ samples are consistently radiogenic in their $^{87}\text{Sr}/^{86}\text{Sr}$ and $^{187}\text{Os}/^{186}\text{Os}$ ratios and have relatively high values of the $\Delta^{207}\text{Pb}$ index¹¹, which is consistent with the association of high $\delta^{18}\text{O}$ with 'enriched' radiogenic-isotope compositions. However, some lower- $\delta^{18}\text{O}$ samples also have enriched radiogenic-isotope compositions, and thus the relationship of oxygen isotopes to radiogenic-isotope enrichment is more complex than that to elemental enrichment (Fig. 1).

Values of $\delta^{18}\text{O}$ for samples examined in this study are not correlated with the abundances of incompatible elements that are sensitive to the extent of melting but are not usually considered to reflect enriched geochemical signatures in MORBs (for example, $\text{Na}_2\text{O}_{(8.0)}$ and TiO_2 ; ref. 22, see ref. 20 for major element compositions). In addition, most of the parameters plotted in Fig. 1 (that is, those other than K_2O , $\text{K}_2\text{O}/\text{TiO}_2$ and La/Ti) are expected to be sensitive only to variations in source composition, provided the total extents of melting of the sources are more than about 1–2% (ref. 23). Although a component of very low degree melt appears to influence the trace-element geochemistry of some MORBs (ref. 24), such melts are believed to be only minor components of erupted lavas. Therefore the small (less than $\sim 0.1\%$) differences in melt-residue oxygen-isotope fractionations expected between high- and low-degree melts of a homogeneous source²⁵ are unlikely to contribute significantly to the observed variations in $\delta^{18}\text{O}$ of erupted lavas. We therefore conclude that the trends in Fig. 1 are unlikely to reflect variable extents of melting of a homogenous source. However, there are few experimental constraints on oxygen-isotope partitioning during such processes, and thus this hypothesis should be re-examined when such data become available. Neither

$\delta^{18}\text{O}$ nor other variables illustrated in Fig. 1 are well correlated with Mg number (the principal index of differentiation in basaltic lavas; Table 1). Furthermore, most elemental indices correlated with $\delta^{18}\text{O}$ in Fig. 1 are insensitive to fractional crystallization. We therefore also conclude that the variations in $\delta^{18}\text{O}$ and relationships of $\delta^{18}\text{O}$ to chemistry in the MORBs studied here do not reflect magmatic differentiation alone. Finally, $\delta^{18}\text{O}$ values are not correlated with H_2O abundances (Table 1), suggesting that they do not reflect the assimilation of water-rich rocks or sediments in the young oceanic crust through which they were erupted or ^{18}O -enrichment during subsolidus hydration of the sampled glasses¹⁷.

We conclude from the discussion above that the variations in $\delta^{18}\text{O}$ observed in the MORB glasses included in this study principally reflect variations in the $\delta^{18}\text{O}$ of their mantle sources. Given the association of elevated $\delta^{18}\text{O}$ with enrichments in elements that are concentrated in high- $\delta^{18}\text{O}$ oceanic crustal rocks and sediments (for example, K, Ba), the trends in Fig. 1 could reflect the presence of variable amounts of ^{18}O -enriched oceanic crustal materials in the sources of the studied lavas. This hypothesis can also explain the lack of a relationship between $\delta^{18}\text{O}$ and H_2O content (Table 1). That is, crustal materials are believed to be strongly dehydrated during subduction²⁶, so their addition to the upper mantle could enrich it in elements that are concentrated in the oceanic crust but not efficiently removed during subduction without strong enrichment in H_2O . We examined this hypothesis by comparing our data with a model in which mantle peridotite that is normal in $\delta^{18}\text{O}$ and relatively poor in incompatible elements mixes with subducted, dehydrated oceanic upper-crustal rocks and sediments. According to the model, this mixed source then melts by a constant amount to generate MORBs (solid curves, Fig. 1; the dashed curves in Fig. 1e and f show the results of an alternative model in which enrichments are introduced as separate domains of eclogite; see Methods for details). We note that mixing between mantle peridotites and oceanic, lower-crustal gabbros would probably produce trends of decreasing $\delta^{18}\text{O}$ with increasing enrichment and would therefore be inconsistent with our observations. However, the contrasts in the values of $\delta^{18}\text{O}$ and of most incompatible-element concentrations between mantle peridotites and oceanic lower-crustal gabbros^{17,27} are subtle, and therefore the trends in Fig. 1 would also be consistent with models in which the recycled materials contain both upper-oceanic crust and a component of gabbro from the lower-oceanic crust.

The observed trends in Fig. 1 are well described by the crustal recycling model described above and are consistent with the presence of 0–7 wt% of subducted upper-crustal material in the sources of MORB magmas ranging from NMORB to transitional EMORB. We conclude that recycling of subducted oceanic crust plausibly controls variations in $\delta^{18}\text{O}$ and correlated aspects of the incompatible-element geochemistry of these MORBs. The slope of

Table 2 Compositions of components for crustal-recycling model

	Depleted mantle*	Recycled sediment†	Recycled, altered MORB‡	Bulk recycled upper crust‡	DII (solid/melt) peridotite	D (solid/melt) eclogite
wt%						
K_2O	0.005	2.04	0.40	0.48	0.0053	0.0041
TiO_2	0.199	0.62	1.35	1.31	0.086	0.44
H_2O	0.026	0.2 (7.3)	0.2 (2.0)	0.2	0.012	0.012
p.p.m.						
Ba	0.56	776	30	67	0.0003	0.00059
U	0.004	1.68	0.14	0.22	0.0007	0.00195
Th	0.011	6.91	0.19	0.53	0.0077	0.0015
La	0.15	28.8	3.0	4.3	0.0061	0.028
Sm	0.24	5.78	2.6	2.8	0.0287	0.312
Sr	12	327	80	92	0.0183	0.08
$\delta^{18}\text{O}\text{‰}$	5.4	20	10	10.5		

* Defined such that the composition of a 10% batch partial melt will have a composition near the low- $\delta^{18}\text{O}$ end of the trends in Fig. 1.

† Water content after dehydration assumed based on K/H versus La/Sm systematics in MORB from ref. 15. Water content before subduction given in italics.

‡ Assuming 5% sediment, 95% altered oceanic crust.

§ $\delta^{18}\text{O}$ of listed source component, estimated for crustal components using data from refs 17, 18, and references supplied in the Supplementary Information.

|| D is the distribution coefficient defined as the ratio of concentrations (by weight) between two coexisting phases (for example, solid/melt) at equilibrium.

All other elemental abundances and distribution coefficients are based on data in references provided in the Supplementary Information and ref. 15.

the trend for our data in Fig. 1e (and perhaps Fig. 1f) agrees more closely with the model in which the enriched material is present as mafic domains (dashed curves) than with the model that assumes enrichment to be a compositional component of a mineralogically homogeneous peridotite (solid curves), consistent with previous suggestions of such a distribution of recycled crust in the sources of MORBs (see ref. 13). We note that even the depleted ends of the trends in Fig. 1 need not be entirely free of recycled crust; if not, the quantitative modelling we have presented would underestimate the importance of recycled materials in the incompatible-element and oxygen-isotope geochemistry of the upper-mantle sources of MORBs.

Given estimates of the average elemental composition of NMORB (ref. 28), our model suggests that their upper-mantle sources contain an average of ~2% recycled upper-crustal oceanic rocks and sediments. If this estimate represents a steady-state condition⁵, and if we assume that the MORB reservoir corresponds to the upper 660 km of the mantle⁵ and that residence times of incompatible elements within it are ~1 Ga (ref. 5), then our result would require that the long-term rate of addition of upper-crustal rocks and sediments to the sources of MORB is approximately 6 km³ per year. This estimate is close to the current rate at which upper oceanic crust is both produced and subducted (~1/3 of the rate of total oceanic crust production and consumption, given the typical stratigraphy of the oceanic crust²⁹), suggesting that the quantities of upper-crustal material that we call upon as an agent of chemical heterogeneity in the sources of MORBs are reasonable. The abundances of incompatible elements used in our models for the depleted mantle and recycled upper-crustal components (Table 2) and the composition of average NMORB (ref. 28) suggest that, on average, approximately half of the total K, Ba and U, and 10–30% of the Sr and rare-earth elements in the upper mantle sources of NMORB, are introduced by subducted oceanic crust. Our interpretation suggests that subduction of oceanic crust into the sources of MORBs has shifted the average $\delta^{18}\text{O}$ of those sources by ~0.1‰ with respect to depleted upper-mantle peridotites. This result constrains models of the oxygen-isotope budget of the upper

mantle–crust–hydrosphere system, although detailed exploration of such models is beyond our scope here.

The fact that the data from the Atlantic, Indian and Pacific MORB samples studied here all follow the same trends in Fig. 1 suggests that distinctions among the major ocean basins in other geochemical characteristics (see ref. 11) do not apply to variations in $\delta^{18}\text{O}$ and incompatible-element abundances among the sources of MORB. Instead, our results suggest that compositional heterogeneities contributing to the variability of oxygen-isotope ratios and correlated incompatible-element characteristics in the sources of MORBs are similar in origin and magnitude over a global length scale and may be a background heterogeneity onto which other types of enrichment are superimposed. This interpretation supports previous conclusions that sources of MORBs contain variable amounts of globally disseminated, enriched components and suggests that such components are principally derived from recycling of the oceanic crust^{7,16} rather than from intramantle differentiation^{8–10}. □

Methods

Measurements of $\delta^{18}\text{O}$ were made by laser fluorination at the California Institute of Technology using a 50-watt CO_2 laser, BrF_3 as the reagent, and a gas-handling apparatus similar to previous designs (see Supplementary Information); all gases were analysed on a Finnigan 251 gas-source isotope-ratio mass spectrometer. Data were standardized using several silicate standards (SCO-1 and GMT-2, see Supplementary Information; and AH95-22, a basalt glass standard used at Caltech); data gathered on separate days were normalized to a common value for these standards (the average correction for any given day was 0.06). After normalization, replicate analyses of AH95-22 glass standard had a standard deviation of ± 0.03 (n = 27); the standard deviations for replicate analyses of unknowns also averaged ± 0.03 ‰ and are listed for individual samples in Table 1. The uncertainty in the mean for each sample (that is, $1/\sqrt{n}$) averaged ± 0.02 ‰ (also listed in Table 1 for unknowns). Measurements of water content were made by Fourier transform infrared spectroscopy at Caltech, with methods referenced in the Supplementary Information.

The crustal mixing model (curves in Fig. 1) assumes a depleted mantle source (defined to have a composition consistent with generating lavas at the low- $\delta^{18}\text{O}$ ends of the trends in Fig. 1; see Table 2) into which oceanic upper-crustal material (assumed to be a mixture of 5% average sea-floor sediment and 95% average altered sea-floor basalt; Table 2) is mixed. Note that oceanic sediments and altered crustal rocks are assumed to be significantly dehydrated but otherwise compositionally similar to existing seafloor materials. This source then undergoes 10% batch partial melting using bulk partition

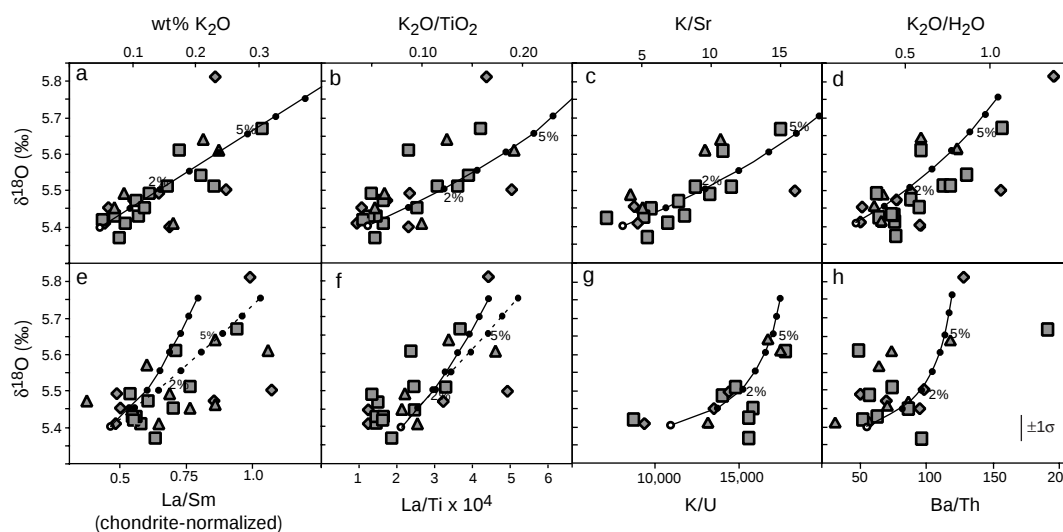


Figure 1 $\delta^{18}\text{O}$ values versus minor- and trace-element contents and ratios in MORB glasses. **a–h**, Data are from ref. 20 and references therein or in the Supplementary Information; water contents (**d**) from this study. All abundance ratios are on a weight basis. Solid curves are calculated compositions of basaltic melts formed by 10% batch melting of depleted mantle containing a variable component of recycled upper-oceanic crust and sediment. White circles mark the locations of melts of model depleted mantle alone; black circles show the locations of partial melts of depleted mantle containing variable amounts of recycled crust in increments of 1%; compositions of melts derived from sources with

2% and 5% recycled crust are labelled. Dashed curves in **e** and **f** are calculated compositions of basaltic melts formed by mixing of 10% batch partial melts of depleted peridotite with 10% batch partial melts of eclogite having the composition of recycled upper-oceanic crust. Dashed curves are omitted from panels other than **e** and **f** because of their close correspondence to the solid curves. See Table 2, Methods, and text for details and discussion. Symbols discriminate samples by major ocean basin: squares, Pacific; diamonds, Atlantic; triangles, Indian. The illustrative error bar is the typical standard deviation ($\pm 1\sigma$) of replicate measurements of unknowns (Table 1).

coefficients listed in Table 2; resulting melt compositions are shown in Fig. 1 as circles joined by solid curves (the white circle is the model melt of the assumed depleted source; the black circles are for melts of sources to which 1–7 wt% of the crustal component has been added in 1% increments). Sediments and altered crustal rocks vary considerably in their composition (particularly in their Ba/Th ratios; Fig. 1h) and therefore our model curves are best regarded as averages about which the range of melts of natural mantle–crust mixtures might scatter. The calculated solid curves in Fig. 1 assume that partition coefficients for trace elements and oxygen isotope fractionation between melt and residue during melting are independent of enrichment. This approximation is likely to be incorrect for the La/Sm and La/Ti ratios if enriched material is present as mafic domains because Sm and Ti are relatively compatible in garnet (a major constituent of high-pressure mafic rocks). We have therefore shown an alternative model in Fig. 1e and f in which MORBs are mixtures of 10% batch partial melts of depleted peridotite and 10% batch partial melts of eclogite having a composition equal to our recycled-crust model component; eclogite is assumed to be 50% garnet and 50% clinopyroxene. Eclogite probably melts to a different extent from peridotite at a given set of conditions. The direction and extent of these differences are expected to vary with temperature, pressure and water fugacity; rather than specifying these variables we have adopted the simplifying assumption that eclogite and peridotite components melt to comparable extents. The slopes of curves calculated by this model are not significantly changed by adopting other reasonable assumptions about the melting behaviour of eclogite, although for some models the amount of recycled material in the source may be significantly different from the fraction of melt derived from that material. This alternative differs little from the first model for indices other than La/Sm and La/Ti as a result of the subtle differences between melt–peridotite and melt–eclogite distribution coefficients for other elements plotted in Fig. 1. We therefore omit the results of this alternative calculation from panels other than Fig. 1e and f. We have assumed that oxygen–isotope fractionations during melting are constant (that is, variations in the $\delta^{18}\text{O}$ of melts are directly proportional to variations in their sources). Data sources for our model calculations are referenced in the Supplementary Information.

Received 12 July; accepted 9 December 1999.

- Wood, J. A. et al. in *Basaltic Volcanism on the Terrestrial Planets* 631–699 (Pergamon, New York, 1981).
- Tatsumoto, M. Genetic relations of oceanic basalts as indicated by lead isotopes. *Science* **153**, 1094–1101 (1966).
- Kay, R., Hubbard, N. J. & Gast, P. W. Chemical characteristics and origin of oceanic ridge volcanic rocks. *J. Geophys. Res.* **75**, 1585–1614 (1970).
- O’Nions, R. K., Hamilton, P. J. & Evenson, A. M. Variations in $^{143}\text{Nd}/^{144}\text{Nd}$ and $^{87}\text{Sr}/^{86}\text{Sr}$ ratios in oceanic basalts. *Earth Planet. Sci. Lett.* **34**, 13–22 (1977).
- Allègre, C. J. Isotope geodynamics. *Earth Planet. Sci. Lett.* **86**, 175–203 (1987).
- Zindler, A. & Hart, S. Chemical geodynamics. *Annu. Rev. Earth Planet. Sci.* **14**, 493–571 (1986).
- Armstrong, R. L. A model for the evolution of strontium and lead isotopes in a dynamic Earth. *Rev. Geophys.* **6**, 175 (1968).
- DePaolo, D. J. & Wasserburg, G. J. Inferences about magma sources and mantle structure from variations of $^{143}\text{Nd}/^{144}\text{Nd}$. *Geophys. Res. Lett.* **3**, 743–746 (1976).
- Green, D. H. Composition of basaltic magmas as indicators of conditions of origin: application to oceanic volcanism. *Phil. Trans. R. Soc. Lond. A* **268**, 707–725 (1971).
- McKenzie, D. & O’Nions, K. Mantle reservoirs and ocean island basalts. *Nature* **301**, 229–231 (1983).
- Hart, S. R. A large-scale isotope anomaly in the southern hemisphere mantle. *Nature* **309**, 753–757 (1984).
- Schilling, J. G. Iceland mantle plume: geochemical study of Reykjanes Ridge. *Nature* **242**, 565–571 (1973).
- Zindler, A., Staudigel, H. & Batiza, R. Isotope and trace element geochemistry of young Pacific seamounts: Implications for the scale of upper mantle heterogeneity. *Earth Planet. Sci. Lett.* **70**, 175–195 (1984).
- Sun, S. S., Nesbitt, R. W. & Sharaskin, A. U. Geochemical characteristics of mid-ocean ridge basalts. *Earth Planet. Sci. Lett.* **44**, 119–138 (1979).
- Michael, P. Regionally distinctive sources of depleted MORB: Evidence from trace elements and H_2O . *Earth Planet. Sci. Lett.* **131**, 301–320 (1995).
- Hofmann, A. W. Mantle plumes from ancient oceanic crust. *Earth Planet. Sci. Lett.* **57**, 421–436 (1982).
- Muehlenbachs, K. in *Stable Isotopes in High Temperature Geological Processes* (eds Valley, J. W., Taylor, H. P. & O’Neil, J. R.) *Reviews in Mineralogy* Vol. 16, 425–444 (Mineralogical Society of America, Washington, DC, 1986).
- Arthur, M. A., Anderson, T. F. & Kaplan, I. R. in *Stable Isotopes in Sedimentary Geology, SEPM Short Course* Vol. 10, 1–151 (SEPM, Tulsa, Oklahoma, 1983).
- Harmon, R. S. & Hoefs, J. Oxygen isotope heterogeneity of the mantle deduced from global ^{18}O systematics of basalts from different tectonic settings. *Contrib. Mineral. Petrol.* **120**, 95–114 (1995).
- Schiano, P., Birck, J. -L. & Allègre, C. J. Osmium–strontium–neodymium–lead isotopic covariations in mid-ocean ridge basalt glasses and the heterogeneity of the upper mantle. *Earth Planet. Sci. Lett.* **150**, 363–379 (1997).
- Mahoney, J., Le Roex, A. P., Peng, Z., Fisher, R. L. & Natland, J. H. Southwestern limits of Indian Ocean ridge mantle and the origin of low $^{206}\text{Pb}/^{204}\text{Pb}$ mid-ocean ridge basalt: isotope systematics of the central Southwest Indian Ridge (17 degrees – 50 degrees E). *J. Geophys. Res.* **B97**, 19771–19790 (1992).
- Langmuir, C. H., Klein, E. M. & Plank, T. in *Mantle Flow and Melt Generation at Mid-Ocean Ridges* (eds Morgan, J. P., Blackman, D. K. & Sinton, J. M.) 183–280 (American Geophysical Union, Washington, DC, 1992).
- Shaw, D. M. Trace element behaviour during anatexis. *Geochim. Cosmochim. Acta* **34**, 237–243 (1970).
- McKenzie, D. & O’Nions, R. K. Partial melt distributions from inversion of rare earth element concentrations. *J. Petrol.* **32**, 1021–1091 (1991).
- Matthews, A., Stolper, E. M., Eiler, J. M. & Epstein, S. Oxygen isotope fractionation among melts, minerals, and rocks. *Min. Mag.* **62A**, 971–972 (1998).
- Poli, S. & Schmidt, M. W. H_2O transport and release in subduction zones: Experimental constraints on basaltic and andesitic systems. *J. Geophys. Res.* **100**, 22299–22314 (1995).

- Pallister, J. S. & Knight, R. J. Rare-earth element geochemistry of the Samail ophiolite near Ibra, Oman. *J. Geophys. Res.* **86**, 2673–2697 (1981).
- Sun, S. S. & McDonough, W. F. in *Magmatism in the Ocean Basins* (eds Saunders, A. D. & Norry, M. J.) 313–345 (Geological Society of London Special Publication 2, 1989).
- Crisp, J. A. Rates of magma emplacement and volcanic activity. *J. Volcanol. Geotherm. Res.* **20**, 177–211 (1984).

Supplementary information is available on Nature’s World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial offices of Nature.

Acknowledgements

We thank P. Michael; we also thank D. Anderson, P. Asimow, A. Halliday and C. Langmuir for comments on the manuscript. We thank the Chevron Corporation for donation of the mass spectrometer used for isotopic measurements.

Correspondence and requests for materials should be addressed to J.M.E. (e-mail: eiler@gps.caltech.edu).

Quality of the fossil record through time

M. J. Benton*, M. A. Wills*† & R. Hitchin*

* Department of Earth Sciences, University of Bristol, Bristol BS8 1RJ, UK

† Oxford University Museum of Natural History, Parks Road, Oxford OX1 3PW, UK

Does the fossil record present a true picture of the history of life^{1–3}, or should it be viewed with caution^{4–6}? Raup⁵ argued that plots of the diversification of life² were an illustration of bias: the older the rocks, the less we know. The debate was partially resolved by the observation⁷ that different data sets gave similar patterns of rising diversity through time. Here we show that new assessment methods, in which the order of fossils in the rocks (stratigraphy) is compared with the order inherent in evolutionary trees (phylogeny), provide a more convincing analytical tool: stratigraphy and phylogeny offer independent data on history. Assessments of congruence between stratigraphy and phylogeny for a sample of 1,000 published phylogenies show no evidence of diminution of quality backwards in time. Ancient rocks clearly preserve less information, on average, than more recent rocks. However, if scaled to the stratigraphic level of the stage and the taxonomic level of the family, the past 540 million years of the fossil record provide uniformly good documentation of the life of the past.

The reduction in quality of the fossil record backwards in time seems self-evident. Fossils in ancient rocks are more likely to have been eroded, crushed, melted, subducted, not collected or misunderstood than younger fossils⁵. The demonstration⁷ that similar patterns of diversification were found from different data sets seemed to indicate that, at least when viewed on a broad scale, the fossil record did correctly document the history of life. Indeed, palaeobiologists subsequently used the fossil record as a literal source of data on the history of the diversification of life^{8,9} and of mass extinctions^{10,11} without applying any correction factors for possible time-related bias. However, the different data sets compared⁷ were all subject to the same geological biasing factors, and that study did not use independent data to demonstrate that older parts of the fossil record could be trusted. This lingering doubt about the long-term quality of the fossil record has resurfaced in debates about the timing of origin of major groups of organisms: molecular studies suggest that the Metazoa (animals)¹² and modern bird and mammal orders^{13,14} apparently originated much earlier

Table 1 Time-related partitions of the dataset of 1,000 cladograms

Dataset partition	Number of trees	Mean tree size	s.e.	Mean range of origins (Myr)	s.e.
1. All trees	1,000	10.359	0.196	126.80	4.19
2. Origins entirely Palaeozoic	156	8.468	0.345	73.17	5.99
3. Origins a mixture of Palaeozoic and post-Palaeozoic dates	261	11.368	0.417	266.36	10.25
4. Origins entirely Mesozoic	106	9.075	0.425	82.70	5.73
5. Origins a mixture of Mesozoic and Cenozoic dates	243	12.086	0.471	123.85	4.12
6. Origins entirely Cenozoic	234	9.282	0.352	29.92	2.29

Numbers of trees in each partition, the mean tree size and the standard error of the range of tree sizes are shown. In these samples, 95% of all trees fall in the range of the mean tree size ± 2 standard errors (s.e.) thus 10.359 ± 0.392 , for all trees.

Table 2 Results for comparisons of cladograms, with stratigraphic data

Dataset partition	SCI Mean	s.e.	RCI Mean	s.e.	GER Mean	s.e.
1	0.551	0.008	31.130	4.587	0.562	0.010
2	0.618	0.021	62.064	3.446	0.529	0.031
3	0.493	0.014	16.247	14.442	0.515	0.018
4	0.601	0.026	44.566	4.368	0.636	0.033
5	0.496	0.013	19.033	6.098	0.627	0.018
6	0.604	0.018	33.604	8.459	0.532	0.024

Data shown according to the stratigraphic consistency index, SCI, the relative completeness index, RCI and the gap excess ratio, GER. Mean values for each metric are given, together with the standard error (s.e.). Dataset partitions are defined in Table 1.

than was expected from the fossil evidence. These new results could be mistaken^{15,16}, or they could indicate long episodes of missing fossil record^{12–14}. However, there is no doubt that older parts of the fossil record cannot ever be as well known as more recent parts. The question is whether the older fossil records are adequate to recount important events in the history of life.

One solution is to compare independent sources of data. New methods for inferring phylogenies, such as cladistics applied to the morphology of fossil and living taxa, and molecular sequencing techniques, provide representations of evolutionary trees that are essentially independent of stratigraphic data^{3,17–19}. Hence, it is possible to compare age data from fossils in the rocks with clade data from molecular and morphological trees using a range of congruence metrics, the stratigraphic consistency index (SCI), the relative completeness index (RCI) and the gap excess ratio (GER) (see Methods).

In our study, we compiled a database of 1,000 morphological and molecular cladograms from the literature (see Methods). The cladograms were sorted into those in which all lineages arose in the Palaeozoic, Mesozoic or Cenozoic, as well as mixed cases with Palaeozoic and post-Palaeozoic origins and mixed Mesozoic/Cenozoic origins. Sample sizes and tree sizes were relatively uniform across all five temporal divisions (Table 1). Remarkably, the relative congruence of age and clade data for all five time divisions showed no clear increase from most ancient to most recent (Table 2; Fig. 1).

In more detail, SCI values for all time partitions of the data set are very similar, with mean values in the range 0.493–0.618, and with most lying close to the mean value for the whole dataset of 0.551. Indeed, the highest mean SCI value overall, 0.618, is for the Palaeozoic-only values, and the lowest mean values (0.493, 0.496) are for the mixed Palaeozoic/Mesozoic and Mesozoic/Cenozoic partitions (Table 2, Fig. 1). The RCI measures showed greater variation, but no clear relationship to time (Table 2; Fig. 1). Mean values range widely, from 11.362 to 62.064, around the mean for the whole dataset of 31.130. The highest value is for the Palaeozoic partitions, which is not surprising as the RCI depends on a measure of total known range (see Methods), and for clades arising in the Palaeozoic, known ranges may potentially extend to the present day. The GER measures show little evidence of time bias (Table 2; Fig. 1). Mean values range from 0.491 to 0.636, but most lie close to the

whole-sample mean of 0.562. The mean value for Palaeozoic cladograms (0.529) is not significantly different from that for the Cenozoic (0.532).

The expected result, that the fossil record of stratigraphic first appearance should become worse with increasing age⁵ is not confirmed. These counter-intuitive results might be the result of temporal biases in the data set^{19–23}, such as (1) variable cladogram quality, (2) differing major taxonomic groups, (3) differing approaches to cladogram construction, (4) the quality of age dates and stratigraphy, (5) the total age span of origins, (6) the taxonomic level of terminal taxa, (7) tree size and (8) tree balance.

The first four of these possible sources of error imply highly unlikely assertions, such as that most Palaeozoic cladograms are markedly better than post-Palaeozoic (1), that the Palaeozoic groups are more amenable to cladistic analysis than the post-Palaeozoic (2), that systematists working on groups with Palaeozoic origins use better techniques than those working on post-Palaeozoic groups (3) and that Palaeozoic age dates and stratigraphy are better than post-Palaeozoic (4).

It could be argued that the first of these problems, variable cladogram quality, makes our results meaningless. The assertion would simply be that cladograms are so full of error that there is no reason to expect congruence with any stratigraphic signal in a large-scale analysis such as ours. There is no absolute measure of cladogram quality. However, cladograms tend to match stratigraphy better than predicted by null models^{19–21,24–26}, and there is evidence that published morphological and molecular phylogenetic trees are probably generally close to the truth.

The remaining four possible biasing factors all turn out to be differently distributed in the current data set. The total span of time occupied by the origins of the groups in an assessed cladogram (5)

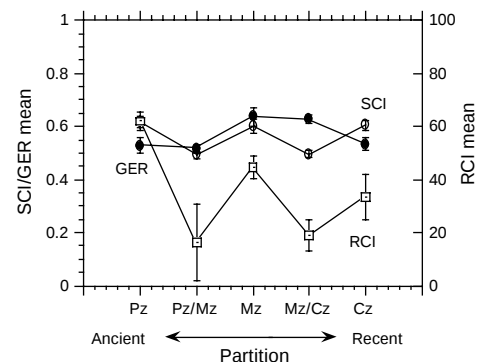


Figure 1 No change in fossil record quality through time. Mean scores of the stratigraphic consistency index (SCI; open circles), the relative completeness index (RCI; open-squares) and the gap excess ratio (GER; filled circles) for five geological time partitions of the data set of 1,000 cladograms. Pz, cladograms with origins solely in the Palaeozoic; Pz/Mz, cladograms with origins spanning the Palaeozoic and Mesozoic; Mz, cladograms with origins solely in the Mesozoic; Mz/Cz, cladograms with origins spanning the Mesozoic and Cenozoic; Cz, cladograms with origins solely in the Cenozoic.

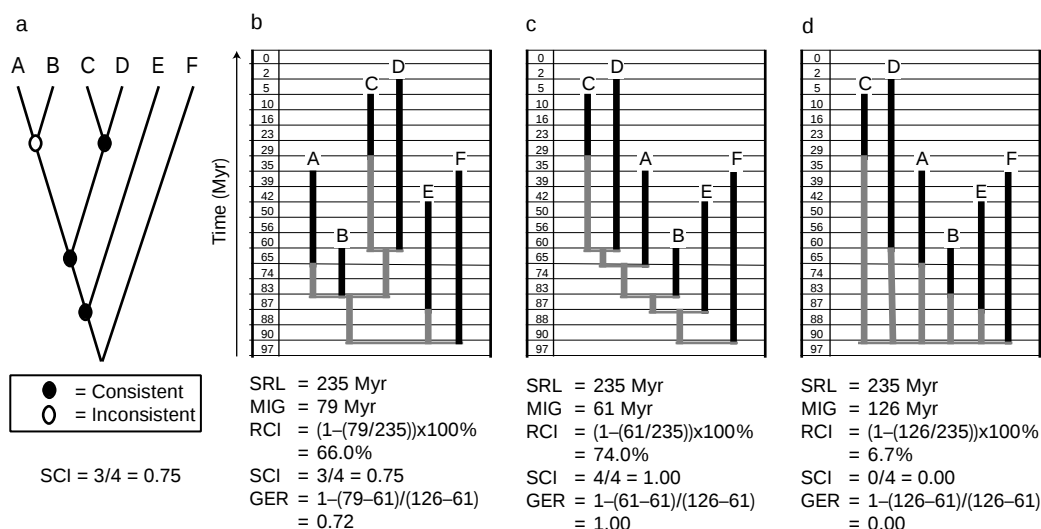


Figure 2 Calculation of the three congruence metrics for age versus clade comparisons. SCI is the ratio of consistent to inconsistent nodes in a cladogram. RCI is $RCI = 1[\Sigma MIG/\Sigma SRL] \times 100\%$ where MIG is minimum implied gap, or ghost range, and SRL is standard range length, the known fossil record. GER is $GER = 1(MIG - G_{min})/(G_{max} - G_{min})$, where G_{min} is the minimum possible sum of ghost ranges and G_{max} the maximum, for any given distribution of origination dates. **a**, The

observed tree with SCI calculated according to the distribution of ranges in **b**. **b**, the observed tree and observed distribution of stratigraphic range data, yielding an RCI of 66.0%. GER is derived from G_{min} and G_{max} values calculated in **c** and **d**. **c**, The stratigraphic ranges from **b** rearranged on a pectinate tree to yield the smallest possible MIG or G_{min} . **d**, the stratigraphic ranges from **b** rearranged on a pectinate tree to yield the largest possible MIG or G_{max} .

can affect all three metrics^{19,20}. However, the situation is complex: all other factors being equal, long origin spans give high SCI values, but the exact opposite applies to the RCI (refs 20, 21). The taxonomic level of terminal taxa (6) might impose a bias: cladograms consisting largely of species and genera tend to fit stratigraphy less well than those based on higher taxa such as families and orders^{19,20}. However, there is no significant variation in taxonomic level through time in our dataset. Tree size (7) has a biasing effect on the SCI: large trees should have lower SCI values than small trees²², but the relationship of RCI and GER values to tree size is unclear²¹. In the present data set, tree size is uniform across all partitions of the data (Table 1). The final potentially biasing variable is tree balance (8); the variation in overall tree shape from symmetrical, or fully balanced, to pectinate (comb-like), or fully imbalanced. Theoretically, the SCI is negatively correlated with tree imbalance, so that imbalanced trees have higher SCI values than balanced trees²². However, no significant relationship between RCI or GER and tree balance has been found in real data^{20,21}. In any case, tree balance is uniform through all time divisions in our data set.

An alternative assertion could be that the age/clade metrics cannot in fact detect major gaps in the fossil record. Perhaps the fossil record does deteriorate seriously backwards in time, but the preserved parts of the fossil record are equally complete through time. For example, a sequence of Cretaceous fossils with ammonites might be no more complete than a Silurian sequence with graptolites, some 300 million years older. The cladogram metrics could very well give equivalent values. However, perhaps the good Silurian record represents only a tiny fraction of the total record, a much smaller proportion than is represented by the Cretaceous ammonites. In other words, the older record is substantially worse, and our metrics might fail to detect it. This would mean that the fossil record is itself a biased sample of past life. However, soft-bodied organisms are equally poorly known from rocks of all ages, for example. The age/clade metrics circumvent the problem to a certain extent, as the phylogenies include preservable and non-preserved taxa alike. Further tests of the fossil record/true record issue are

required. Our results are confirmed by recent studies that consider the preservation potential of different fossil groups³⁰.

Our finding that the fossil record does not diminish in quality with time seems at first impossible, because it is a demonstrable fact that geological activity destroys ancient rocks and ancient fossils⁵. We should consider the scaling of observations in terms of both stratigraphy and taxonomy. Experience shows that major changes in the dating of fossils do not occur at the level of geological systems or stages, but at the finer divisions of substages and zones. Likewise, orders and families are often relatively stable, while new discoveries constantly alter the definitions of genera and species of fossils. The stability of longer time intervals and larger taxonomic categories perhaps reflects an adequate (if incomplete) fossil record. However, global studies of diversification at species and zonal levels would generally be meaningless because the incompleteness of more ancient parts of the fossil record renders it inadequate in most cases for such studies. It is important to distinguish between 'completeness' and 'adequacy'²⁷. Early parts of the fossil record are clearly incomplete, but they can be regarded as adequate to illustrate the broad patterns of the history of life. □

Methods

Data base

The data set consisted of 1,000 cladograms, including one cladogram of 'all life', 33 cladograms of plants, nine cladograms of coelenterates, one cladogram of molluscs, 179 cladograms of arthropods, 14 cladograms of brachiopods, one cladogram each of bryozoans and graptolites, 60 cladograms of echinoderms, 34 cladograms of basal deuterostomes including calcichordates, 157 cladograms of fishes and 510 cladograms of tetrapods, including 26 of amphibians, 203 of reptiles, 8 of birds and 269 of mammals, extracted piecemeal from many sources. The *Fossil Record 2* (ref. 28) was the major source of stratigraphic data for dates of origin of families and suprafamilial taxa. Some cladograms included genera and species, and their dates of origin were generally determined from data provided in the paper that presented the cladogram. Origins and extinctions of clades were assessed to the level of the stratigraphic stage (mean duration of the 79 time units used for the Phanerozoic is 6.8 Myr). Geological dates for these stages (in Myr) were taken from one source²⁹. The data set was divided into stratigraphic divisions listed in Table 1, and the entire data set is available as Supplementary Information.

Age versus clade congruence

Three measures were used to assess age versus clade congruence (Fig. 2): the stratigraphic consistency index, SCI (ref. 24), the relative completeness index, RCI (ref. 25) and the gap excess ratio, GER (ref. 21). The SCI is the ratio of consistent to inconsistent nodes in a cladogram, and it can range from 0 to 1.0 in a fully pectinate (unbalanced) tree, but the minimum value lies between 0 and 0.5 in balanced trees^{21,22}. The RCI and GER depend on numerical age estimates of the branching points on a cladogram, and the calculation of 'ghost ranges'. The ghost range²⁶ is the difference in age, or number of stratigraphic intervals, between the oldest known fossils of two sister taxa. The RCI is assessed as the ratio between the sum of ghost ranges to the sum of recorded fossil ranges in any cladogram. The GER focuses solely on the estimated dates of origin of groups, and compares the sum of actual ghost ranges in a cladogram with the theoretical minimum and maximum ghost ranges if the various branches in the cladogram are rearranged. The metrics were calculated using the software 'Ghosts 2.4', developed by M.A.W., which assesses all three metrics (SCI, RCI, and GER) for individual cladograms, or for large batches of cladograms (available from <http://palaeo.gly.bris.ac.uk/cladestrat/cladestrat.html>).

Received 7 June; accepted 15 November 1999.

1. Simpson, G. G. *Tempo and Mode in Evolution* (Columbia Univ. Press, New York, 1944).
2. Valentine, J. W. Patterns of taxonomic and ecological structure of the shelf benthos during Phanerozoic time. *Palaeontology* **12**, 684–709 (1969).
3. Smith, A. B. *Systematics and the Fossil Record* (Blackwell, Oxford, 1994).
4. Hennig, W. *Phylogenetic Systematics* (Univ. of Illinois Press, Urbana, 1966).
5. Raup, D. M. Taxonomic diversity during the Phanerozoic. *Science* **177**, 1065–1071 (1972).
6. Patterson, C. Significance of fossils in determining evolutionary relationships. *Annu. Rev. Ecol. Syst.* **12**, 195–223 (1981).
7. Sepkoski, J. J. Jr, Bambach, R. K., Raup, D. M. & Valentine, J. W. Phanerozoic marine diversity and the fossil record. *Nature* **293**, 435–437 (1981).
8. Sepkoski, J. J. Jr A kinetic model of Phanerozoic taxonomic diversity. III. Post-Paleozoic families and mass extinctions. *Paleobiology* **10**, 246–267 (1984).
9. Benton, M. J. Diversification and extinction in the history of life. *Science* **268**, 52–58 (1995).
10. Raup, D. M. & Sepkoski, J. J. Jr Mass extinctions in the marine fossil record. *Science* **215**, 1501–1503 (1982).
11. Raup, D. M. & Sepkoski, J. J. Jr Periodicity of extinctions in the geologic past. *Proc. Natl Acad. Sci. USA* **81**, 801–805 (1984).
12. Wray, G. A., Levinson, J. S. & Shapiro, L. H. Molecular evidence for deep Precambrian divergences among metazoan phyla. *Science* **274**, 568–573 (1996).
13. Cooper, A. & Penny, D. Mass survival of birds across the Cretaceous–Tertiary boundary: molecular evidence. *Science* **275**, 1109–1113 (1997).
14. Kumar, S. & Hedges, S. B. A molecular timescale for vertebrate evolution. *Nature* **392**, 917–920 (1998).
15. Ayala, F. J., Rzhetsky, A. & Ayala, F. J. Origin of metazoan phyla: molecular clocks confirm paleontological estimates. *Proc. Natl Acad. Sci. USA* **95**, 606–611 (1998).
16. Foote, M., Hunter, J. P., Janis, C. M. & Sepkoski, J. J. Jr Evolutionary and preservational constraints on origins of biologic groups: divergence times of eutherian mammals. *Science* **283**, 1310–1314 (1999).
17. Forey, P. L. *et al. Cladistics: A Practical Course in Systematics* (Clarendon, Oxford, 1992).
18. Hillis, D. M., Moritz, C. & Mable, B. K. *Molecular Systematics* 2nd edn (Sinauer, Sunderland, MA, 1996).
19. Benton, M. J. & Hitchin, R. Testing the quality of the fossil record by groups and by major habitats. *Historical Biol.* **12**, 111–157 (1996).
20. Benton, M. J., Hitchin, R. & Wills, M. A. Assessing congruence between cladistic and stratigraphic data. *Syst. Biol.* **48**, 581–596 (1999).
21. Wills, M. A. The gap excess ratio, randomization tests, and the goodness of fit of trees to stratigraphy. *Syst. Biol.* **48**, 559–580 (1999).
22. Siddall, M. E. Stratigraphic consistency and the shape of things. *Syst. Biol.* **45**, 111–115 (1996).
23. Wagner, P. J. In *The Adequacy of the Fossil Record* (eds Donovan, S. K. & Paul, C. R. C.) 165–187 (Wiley, New York, 1998).
24. Huelsenbeck, J. P. Comparing the stratigraphic record to estimates of phylogeny. *Palaeobiology* **20**, 470–483 (1994).
25. Benton, M. J. & Storrs, G. W. Testing the quality of the fossil record: paleontological knowledge is improving. *Geology* **22**, 111–114 (1994).
26. Norell, M. A. in *Extinction and Phylogeny* (eds Novacek, M. J. & Wheeler, Q. D.) 89–118 (Columbia Univ. Press, New York, 1992).
27. Paul, C. R. C. in *The Adequacy of the Fossil Record* (eds Donovan, S. K. & Paul, C. R. C.) 1–22 (Wiley, New York, 1998).
28. Benton, M. J. *The Fossil Record 2* (Chapman & Hall, London, 1993).
29. Harland, W. B. *et al. A Geologic Time Scale 1989* (Cambridge Univ. Press, Cambridge, 1993).
30. Foote, M. & Sepkoski, J. J. Jr Absolute measures of the completeness of the fossil record. *Nature* **398**, 415–417 (1999).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature; and may also be viewed at <http://palaeo.gly.bris.ac.uk/cladestrat/cladestrat.html>.

Acknowledgements

We thank the Leverhulme Trust and NERC for continued funding of our work, and E. Fara, M. Foote and P. N. Pearson for helpful comments on the manuscript.

Correspondence and requests for materials should be addressed to M.J.B. (e-mail: mike.benton@bris.ac.uk).

Ontogeny of orientation flight in the honeybee revealed by harmonic radar

Elizabeth A. Capaldi*, Alan D. Smith†, Juliet L. Osborne‡, Susan E. Fahrbach*, Sarah M. Farris*, Donald R. Reynolds†, Ann S. Edwards†, Andrew Martin‡, Gene E. Robinson*, Guy M. Poppy‡ & Joseph R. Riley†

* Department of Entomology, University of Illinois at Urbana-Champaign, 320 Morrill Hall, 505 South Goodwin Avenue, Urbana, Illinois 61801, USA

† Radar Entomology Unit, Natural Resources Institute, University of Greenwich, Leigh Sinton Road, Malvern, Worcestershire, WR14 1LL, UK

‡ Department of Entomology & Nematology, IACR Rothamsted, Harpenden, Hertfordshire, AL5 2JQ, UK

Cognitive ethology focuses on the study of animals under natural conditions to reveal ecologically adapted modes of learning. But biologists can more easily study what an animal learns than how it learns. For example, honeybees take repeated 'orientation' flights before becoming foragers at about three weeks of age¹. These flights are a prerequisite for successful homing.² Little is known^{2,3} about these flights because orienting bees rapidly fly out of the range of human observation. Using harmonic radar, we show for the first time a striking ontogeny to honeybee orientation flights. With increased experience, bees hold trip duration constant but fly faster, so later trips cover a larger area than earlier trips. In addition, each flight is typically restricted to a narrow sector around the hive. Orientation flights provide honeybees with repeated opportunities to view the hive and landscape features from different viewpoints, suggesting that bees learn the local landscape in a progressive fashion. We also show that these changes in orientation flight are related to the number of previous flights taken instead of chronological age, suggesting a learning process adapted to changes in weather conditions, flower availability and the needs of bee colonies.

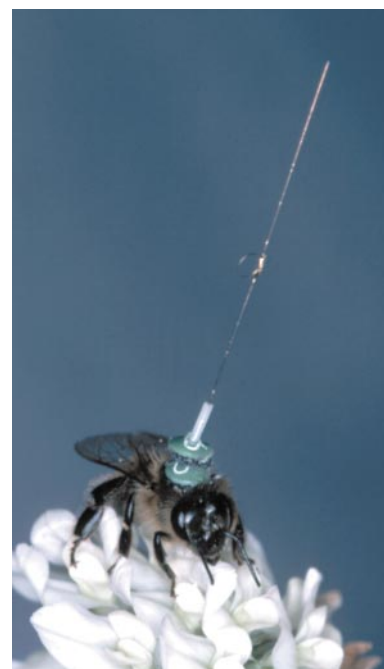


Figure 1 Bee wearing transponder used for harmonic radar tracking of flight. The transponder weighs less than an average load of nectar or pollen.

We began our investigations without harmonic radar, by introducing one-day-old adult worker bees tagged with numbered disks ($n = 125$) into a colony of about 20,000 workers and a queen, and simply watching them as they exited and re-entered the hive. All flight activities of the tagged bees were recorded. We found that no bee became a forager (defined as a bee that returns to the hive with nectar or pollen) without taking at least one orientation flight and that almost all bees took multiple orientation flights before they began to forage. The number of orientation flights taken by bees before foraging was highly variable, ranging from 1 to 18 (mean $\pm$ s.e., 5.6 ± 0.29 flights). Equally variable was the age at which these flights began, ranging from 3 to 14 days (mean $\pm$ s.e., 6.2 ± 0.18 days). In this study the mean age at onset of pollen foraging was 14 ± 2 days.

We examined whether, given this variability, individual orientation flights differ after bees leave the vicinity of the hive. We used harmonic radar^{4,5} to extend the range of our observations in the next study. Tagged one-day-old bees were introduced into a standard hive, and all of their subsequent flights were visually observed and recorded. We were thus able to select bees of known age and flight history for radar tracking. Transponders were attached to bees as they departed from the hive (Fig. 1) and orientation flight tracks were recorded by the radar ($n = 53$). Each bee was tracked only once. We analysed 29 complete (out-and-back) orientation flights taken

by bees of different ages (3–27, median = 6 days), with differing degrees of experience (1–17 orientation flights). For comparison, we also tracked flights ($n = 15$) taken by experienced foragers of unknown age from the same hive. All flights were naturally occurring; that is, all tracked bees had moved spontaneously from inside the hive to the hive entrance before attachment of the transponder.

Most bees fitted with transponders began their orientation flights with a brief period of hovering facing the colony entrance before departure (see Methods), just as those without transponders did and as has been previously reported³. The duration of the orientation flights of bees with and without transponders was not significantly different (with transponders: mean $\pm$ s.e., 331.6 ± 59.2 s ($n = 29$) versus without transponders: 340.1 ± 26.4 s ($n = 219$), Wilcoxon test, $P > 0.05$, n.s.). We therefore concluded that the attachment of transponders did not significantly alter flight behaviour.

Sample flight tracks are shown in Fig. 2, and characteristics of the orientation tracks are summarized in Table 1. We found significant positive correlations between flight number and speed relative to the ground (averaged over the journey), round trip journey distance, maximum range from the hive and the area covered by the convex polygon circumscribing each flight (Table 1, Fig. 3). There was no significant positive correlation between flight number and

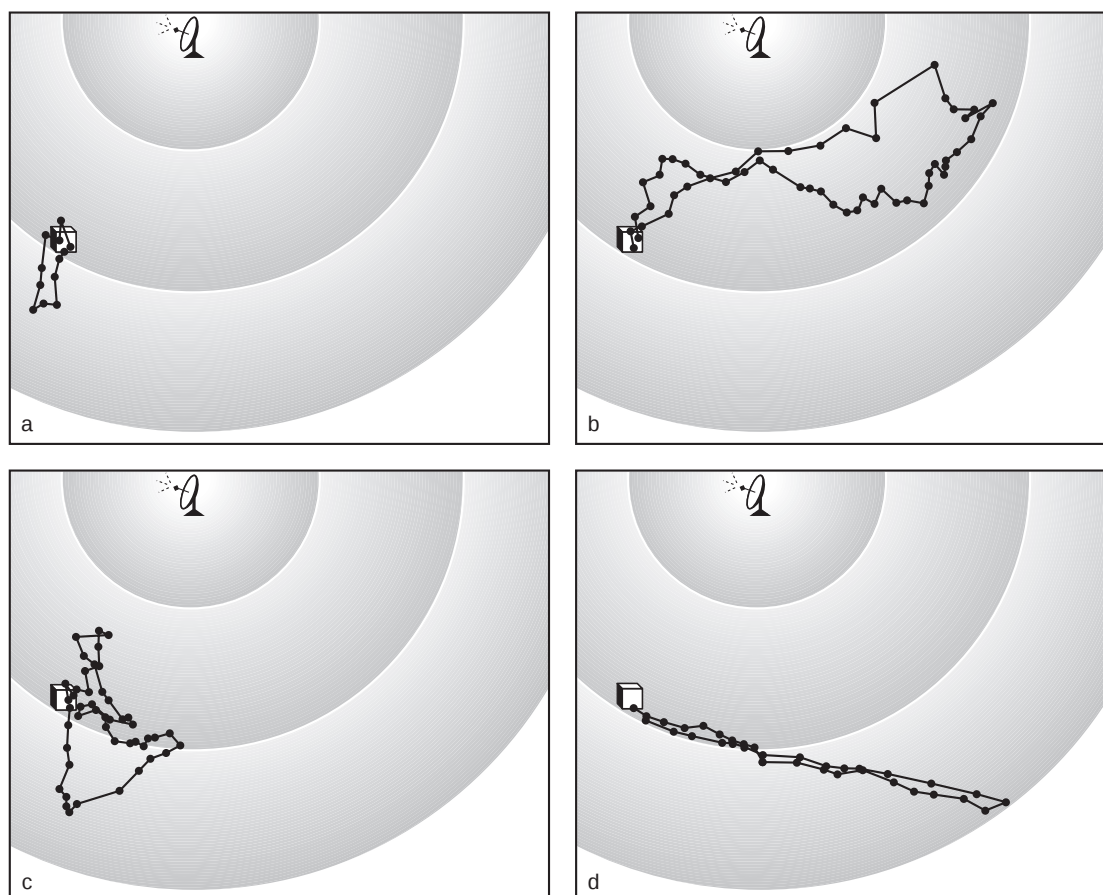


Figure 2 Representative radar tracks from four bees. Orienting bees (**a–c**) were of known age and flight experience. The foraging bee (**d**) was definitively identified as a forager upon her return to the hive with a pollen load. The position of the hive and the harmonic radar are indicated. Shaded area indicates extent of radar coverage. Distance between concentric rings is 116 m. **a**, Orientation flight number 1, age 4 days, duration 228 s. **b**, Orientation flight number 2, age 6 days, duration 1,011 s. **c**, Orientation flight number 14, age 6 days, duration 486 s. **d**, Foraging flight, unknown age forager. Unlike the tracks

shown for orienting bees, this track is not complete, as the bee left the region of radar tracking, re-entering on its homeward journey. Duration, 1,215 s. Comparisons between orientation and foraging flights suggest that both ground speed and flight distance continue to increase as bees mature into foragers. Foraging bees had a significantly faster average ground speed than orienting bees (5.6 ± 1.0 m s⁻¹ versus 3.6 ± 1.4 m s⁻¹) and also flew further and straighter, in agreement with previous reports that honeybees forage over distances of hundreds and even thousands of metres^{10,15}.

flight duration; as the bees gained experience they apparently travelled further by moving faster rather than by staying out longer. Partial correlation analysis indicated that the observed change in ground speed is sufficient to explain the significant correlation between flight number and area (Pearson partial $\rho = 0.14$ for flight number and area with speed as the partial variable; $P = 0.48$). Partial correlation analysis also confirmed that the observed relationship between flight number and ground speed was not caused by changes in wind speed (Pearson partial $\rho = 0.53$ for flight number and ground speed with wind speed as the partial variable; $P = 0.004$). Surprisingly, changes in orientation flights that were correlated with flight number were not also correlated with age (Table 1, Fig. 3), indicating that the rate at which bees acquired orientation experience was variable. The observed experience-related increase in average ground speed during flight is a finding that has not been considered before in studies of orientation behaviour.

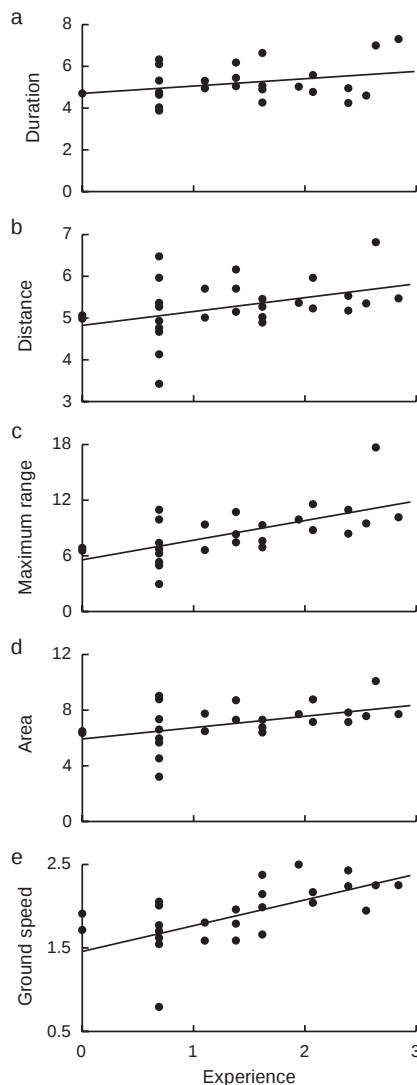


Figure 3 Plots illustrating the correlations between flight experience with the flight attributes. **a–e**, Experience (number of flights), duration (s), distance (m), and area (m²) were natural log transformed. Ground speed (m s^{−1}) and maximum range (m) were square root transformed. Only the correlation shown in **a** (between duration and flight experience) is not significant ($\rho = 0.334$, $P = 0.07$). Nonparametric correlation analyses (Spearman rank order test) on untransformed data matched the conclusions of the parametric correlation analyses.

Table 1 Attributes of honeybee orientation flights tracked with harmonic radar and correlation of flight attributes with age and experience

	Flight Attributes		Age		Experience*	
	Median	Range	$\rho^\dagger$	P	$\rho^\ddagger$	P
Duration (s)	135	48–1,485	0.191	0.32	0.334	0.08
Round trip distance# (m)	189	30–900	−0.234	0.22	0.386	0.04
Maximum range (m)	58	9–314	0.032	0.87	0.619	0.0003
Area covered (m ²)	1418	27–23,997	−0.146	0.45	0.474	0.009
Average ground speed (m s ^{−1})	3.6	0.6–6.2	0.211	0.27	0.599	0.0006

* $N = 29$ flights, one per bee; explanations of measures in Methods.

† Number of orientation flights taken

‡ Pearson correlation coefficient from transformed data, see Fig. 3 for details

Even the long duration flights began with hovering, an indicator of orientation.

See Methods for details

We examined the distribution of the flight tracks around the hive to gain insight into the orienting bee's view of the landscape. Previous studies have shown that bees in some landscapes are able to home after displacement after a single flight in a landscape^{2,6}. Based on these findings, we expected that the tracked bees would explore all quadrants around the hive during a single flight. This rarely occurred. Only 6 out of 29 orientation flights visited all 4 quadrants. In 28 out of 29 tracks, at least 50% of the flight was confined to a single quadrant; thus bees did not typically investigate the entire area surrounding the hive during a single orientation flight. There was a bias towards occupancy of the two quadrants in front of the hive (frequency analysis of radar position fixes per quadrant in each flight: $\chi^2 = 109.5$, degrees of freedom (d.f.) = 3, $P < 0.05$). It has previously been reported that honeybee flight direction is biased toward the position of the sun's azimuth and toward directions in which forage is available³. We studied bees from just one colony, with its entrance continuously facing due east, so we did not test for ecological correlates of this slight directional bias.

Our findings suggest that bees take multiple orientation flights before becoming foragers in order to visit different, and larger, portions of the landscape around the hive. These flights provide them with repeated opportunities to view the hive and its surroundings from different positions, suggesting that bees learn the local landscape in a progressive fashion. Bees navigate using a combination of cues, including the position of the sun and the location of salient landscape features^{7–9}, but it is not known how or whether information about these cues, obtained during sequential flights, is integrated.

It is an extraordinary feat for an animal the size of a honeybee to be able to find a small nest from distances as great as 10 km (ref. 10). Our finding of a progressive process of orientation behaviour begins to demystify this feat for honeybees and may have relevance for other species that face analogous navigational tasks. As these results are incorporated into future models of animal navigation, they will enhance the value of the bee in the fields of cognitive ethology and, eventually, cognitive neuroethology. □

Methods

Bees

Bees were typical of North American and United Kingdom populations of *Apis mellifera*, and were probably a mix of the European subspecies *mellifera* and *ligustica*.

Entrance observations

One-day-old bees were obtained by collecting adults that emerged from incubated frames of capped brood. Numbered, coloured plastic tags were applied to the thorax. For hive entrance observations conducted in Illinois, flight activities of marked bees were recorded from 12:00 to 21:00 until all observed bees had initiated pollen and/or nectar foraging (mean $\pm$ s.e., 14 ± 0.18 days). (Once a bee becomes a forager, she takes orientation flights only when displaced to a new environment or when the visual scene around a familiar location has changed¹¹.) Simulated rain at the hive exit prevented flight activities during the morning hours. Complete records were obtained for 125 bees. In the radar-tracking study at Rothamsted, flight activities of marked bees were recorded from 08:00 to 20:00. Over the five-week course of the study, 625 marked one-day-old bees were introduced into a colony of approximately 10,000 bees. Flights were again prevented by simulated rain

when no observer was present. Flight activities were recorded for all bees until any bee in the cohort was identified as a pollen forager. This procedure ensured that tracked bees were taking orientation flights and were not foragers. Bees for radar tracking were selected on the basis of their flight experience and age. They were captured as they departed from the hive. A queen-marking cage was used to restrain the bee as the transponder was attached to the number tag using double-sided foam tape. The bee was then released from the cage at the height of the hive entrance. When the bee returned to the hive-front, the transponder was removed before she was permitted to re-enter the hive.

Radar tracking

The radar study site was a flat region of farmland (700 m × 900 m) at the Institute of Arable Crops Research, Rothamsted (Harpenden, Hertfordshire, UK). It was planted with cereal crops and was surrounded by flowering crops, hedges, wooded areas and buildings⁵. The radar and transponders have been previously described^{4,5}. The transponder worn by the bees consisted of a 16 mm vertical dipole aerial and a Schottky diode. Two sizes were used, 0.8 mg and 12.0 mg. The larger version weighs less than an average pollen or nectar load¹. Glass or plastic tubing was attached to the bottom of the transponder so that it could be inserted into a small hole drilled into the number tag. This created a base for the transponder, facilitating its attachment to the bee. Tracks were obtained during June and July 1997, on days when the sun or blue sky was visible. Five anemometry stations were set up around the study site, and these recorded wind speed and direction once every 10 s. To obtain a single measure of wind magnitude and direction for each bee flight, the mean wind vector at the midpoint between the beginning and end of the flight path was obtained by interpolation of the vectors recorded at the anemometry stations, averaged over the flight period.

Data analysis

53 tracks were obtained from orienting bees, and 29 of them met the following criteria for analysis: all signals fell within the 700 m radar range and the tracked path started and ended at the hive front. An orientation flight was defined as a flight that began with hovering in front of the hive entrance (23 out of 29); of the remaining 6 tracked bees, most were seen again on subsequent flights preceded by hovering. At the end of her tracked flight, one bee lost her identification tag and could not be studied further. Hovering is a widely accepted indicator of orientation³. Flight speed is equivalent to the mean ground speed for the flight, and was calculated by averaging the point-to-point speed of the moving bee based on 3-s radar sampling intervals. Track segments ≥ 9 s in duration were used for this calculation to match flight-speed sampling intervals to wind-speed information. In cases in which a radar track was divided into segments by a missing signal, the flight-speed measure ignores the time the bee spent in these gaps unless the plotted data indicated that the flight path continued on the same course and speed. Proportion of time spent in gaps was not correlated with age or experience (data not shown), and probably represents short landings on vegetation. The area defined by a tracked flight was calculated using the minimum convex polygon method of home range analysis¹². The *x* and *y* coordinates of each signal were imported into the Antelope spatial analysis software package (FTP://www-biolog.ucsd.edu/research/vehrenbury/programs.html). SAS Institute software was used for all other analyses¹³. All correlation analyses used the Pearson product moment coefficient¹⁴. The data were transformed to satisfy the requirements of this parametric test. Non-parametric tests (Spearman correlation) conducted on the untransformed data yielded identical results (data not shown). Quadrants around the hive were defined by the intersection of two perpendicular lines centred at the hive.

Received 12 August; accepted 10 November 1999.

1. Winston, M. L. *The Biology of the Honey Bee* (Harvard Univ. Press, Cambridge, MA, 1987).
2. Becker, L. Untersuchungen über das Heimfindevermögen der Bienen. *Z. Vergl. Physiol.* **41**, 1–25 (1958).
3. Vollbeh, J. Zur Orientierung junger Honigbienen bei ihrem ersten Orientierungsflug. *Zool. Jb. Allg. Zool. Physiol.* **79**, 33–69 (1975).
4. Riley, J. R. et al. Tracking bees with harmonic radar. *Nature* **379**, 29–30 (1996).
5. Osborne, J. L. et al. A landscape scale study of bumble bee foraging range and constancy, using harmonic radar. *J. Appl. Ecol.* **36**, 519–533 (1999).
6. Capaldi, E. A. & Dyer, F. C. The role of orientation flights on homing performance in honeybees. *J. Exp. Biol.* **202**, 1655–1666 (1999).
7. Collett, T. S. Landmark learning and guidance of insects. *Phil. Trans. R. Soc. Lond. B*, **337**, 295–303 (1992).
8. Lindauer, M. Angeborene und erlernte Komponenten in der Sonnenorientierung der Bienen. *Z. Vergl. Physiol.* **42**, 43–62 (1959).
9. Dyer, F. C. & Dickinson, J. A. How partially experienced bees estimate the sun's course. *Proc. Natl. Acad. Sci. USA* **91**, 4471–4474 (1994).
10. Visscher, P. K. & Seeley, T. D. Foraging strategies of honeybee colonies in a temperate deciduous forest. *Ecology* **6**, 1790–1801 (1982).
11. Wehner, R. in *Handbook of Sensory Physiology* (ed. H. Antrun), Volume VII/6c, 287–616 (Springer, New York, 1981).
12. White, G. C. & Garrott, R. A. *Analysis of Wildlife Radio Tracking Data* (Academic, New York, 1990).
13. *SAS Stat Users Guide* 6. 03 Edition (SAS Institute, Cary, North Carolina, 1988).
14. Zar, J. H. *Biostatistical Analysis* (Prentice Hall, Upper Saddle River, New Jersey, 1996).
15. Ribbands, C. R. *The Behaviour and Social Life of Honeybee*. (Bee Research Association, London, 1953).

Acknowledgements

We thank J. Drew, T. Hallam, J.C. Kuehn, C. Mata and J. Strand for field assistance in Illinois, N. Carreck for beekeeping assistance at Rothamsted, S. Aref for help with data analysis, and F.C. Dyer and the members of the Fahrbach and Robinson laboratories for

reviewing this manuscript. This work was supported by the National Science Foundation, The Research Board of the University of Illinois at Urbana-Champaign, the Biotechnology and Biological Sciences Research Council of the United Kingdom, the British Beekeepers' Association, the European Community Regional Tsetse and Trypanosomiasis Control Programme, the Leverhulme Trust and the United Kingdom Department for International Development Flexibility Fund. E.A.C. was supported by a National Research Service Award from the NIH.

Correspondence and requests for materials should be addressed to E.A.C. (e-mail: ecapaldi@life.uiuc.edu).

Myoglobin-like aerotaxis transducers in Archaea and Bacteria

Shaobin Hou*, Randy W. Larsen†, Dmitri Boudko*, Charles W. Riley*, Ece Karatan‡, Mike Zimmer‡, George W. Ordal‡ & Maqsdul Alam*

* Department of Microbiology, Snyder Hall 207, 2538 The Mall, University of Hawaii, Honolulu, Hawaii 96822, USA

† Department of Chemistry, 2545 The Mall, University of Hawaii, Honolulu, Hawaii 96822, USA

‡ Department of Biochemistry, University of Illinois, 506 S. Mathews Ave, Urbana, Illinois 61801, USA

Haem-containing proteins such as haemoglobin and myoglobin play an essential role in oxygen transport and storage. Comparison of the amino-acid sequences of globins from Bacteria and Eukarya suggests that they share an early common ancestor, even though the proteins perform different functions in these two kingdoms^{1–6}. Until now, no members of the globin family have been found in the third kingdom, Archaea. Recent studies of biological signalling in the Bacteria and Eukarya have revealed a new class of haem-containing proteins that serve as sensors⁷. Until now, no haem-based sensor has been described in the Archaea. Here we report the first myoglobin-like, haem-containing protein in the Archaea, and the first haem-based aerotactic transducer in the Bacteria (termed HemAT-*Hs* for the archaeon *Halobacterium salinarum*, and HemAT-*Bs* for *Bacillus subtilis*). These proteins exhibit spectral properties similar to those of myoglobin and trigger aerotactic responses.

When the *hemAT-Hs* gene was originally cloned, its product was predicted to be a soluble signal transducer⁸ (HemAT-*Hs* was originally named HtB)⁸. HemAT-*Bs* was identified in the *B. subtilis* genome-sequencing project as the product of an open-reading frame encoding a protein with marked similarities to methyl-accepting chemotaxis proteins (MCPs) (HemAT-*Bs* was previously named YhfV)⁹. The predicted translation products of the *hemAT-Hs* and *hemAT-Bs* genes, comprising 489 and 432 residues, respectively, exhibit two striking features: first, their amino termini (residues 1–184 in HemAT-*Hs* and 1–175 in HemAT-*Bs*) display limited homology to myoglobin (Fig. 1a); and, second, residues 222–489 of HemAT-*Hs* and 198–432 of HemAT-*Bs* are 30% identical to the cytoplasmic signalling domain of Tsr (ref. 10), a MCP from *Escherichia coli* (Fig. 1b).

The residues absolutely conserved among all globins are the proximal histidine in the F helix (F8) and phenylalanine in the CD region (CD1)^{11,12}. Highly conserved residues include the distal His in the E helix (E7), Phe in the CD4 region and proline at the beginning of the C helix (C2). Three of these residues (Pro in C2, Phe in CD1, His in F8) are conserved in both HemAT-*Hs* and HemAT-*Bs* (asterisks in Fig. 1a). These features suggested to us that HemATs may be haem-containing proteins that generate signals in response to binding of oxygen.

To identify the prosthetic groups in HemAT-*Hs* and HemAT-*Bs*,

both proteins were expressed in *E. coli* from recombinant vectors. Recombinant HemAT-*Hs* was purified using anion-exchange and gel-filtration chromatography. During SDS–polyacrylamide gel electrophoresis (SDS–PAGE), purified HemAT-*Hs* migrated more slowly than expected from its calculated relative molecular mass (M_r) of 52,800 (52.8K; Fig. 2a, lane 1). This behaviour is consistent with the highly acidic nature of many halophilic proteins¹³. HemAT-*Hs* was purified from *H. salinarum* by metal-affinity and gel-filtration chromatography as a recombinant protein (HemAT_{6xHis}-*Hs*) carrying a carboxyl-terminal six-histidine tag (Fig. 2a, lane 2). HemAT-*Bs* was purified using a combination of ammonium-sulphate precipitation/fractionation and gel-filtration chromatography. As expected, purified HemAT-*Bs* migrated during SDS–PAGE as a 48.7K protein (Fig. 2a, lane 3).

HemAT-*Hs* and HemAT-*Bs* display similar absorption spectra in the near-ultraviolet and visible regions, as is characteristic of oxygen-bound haem proteins. Absorption maxima occur at 406 nm (Soret), 578 nm (α -band) and 538 nm (β -band) for both proteins, compared with corresponding maxima of 418, 581 and 543 nm for horse-heart myoglobin (Fig. 3a). On deoxygenation with sodium dithionite, the Soret bands of the HemAT proteins shift to 425 nm, and the α - and β -bands converge to a broad peak at

555 nm (Fig. 3b). This behaviour is consistent with the formation of a deoxy form of the proteins and is also seen with deoxymyoglobin. If the deoxy forms of HemATs are exposed to atmospheric oxygen, the absorption spectra revert back to those observed with the purified proteins (data not shown). Both the oxy and the deoxy forms of HemATs also react with carbon monoxide (CO), which causes shifts to absorption maxima at 415 nm (Soret), 573 nm (α -band) and 535 nm (β -band) (Fig. 3c). Spectra of both the CO- and O₂-bound forms of HemAT-Bs show the presence of a significant population of deoxy species. We observed that the ligand affinity of HemAT-Bs (and to some extent HemAT-Hs) is dependent on the redox state of the haem before addition of the ligand. Redox cycling through the ferric Fe(III) form of the protein (a small population of which is present on isolation of HemATs) lowers the ligand affinity. This may represent redox-mediated conformational control of oxygen binding. A pyridine haemochrome assay showed that the haem groups of HemATs are of the b-type (data not shown). HemAT_{6×His}-Hs purified from its native host, *H. salinarum*, produced spectral properties essentially identical to those of HemAT-Hs and HemAT-Bs (data not shown), indicating that the six-His tag has no effect on the spectral properties of HemAT-Hs.

To investigate the role of HemATs in signal transduction, we

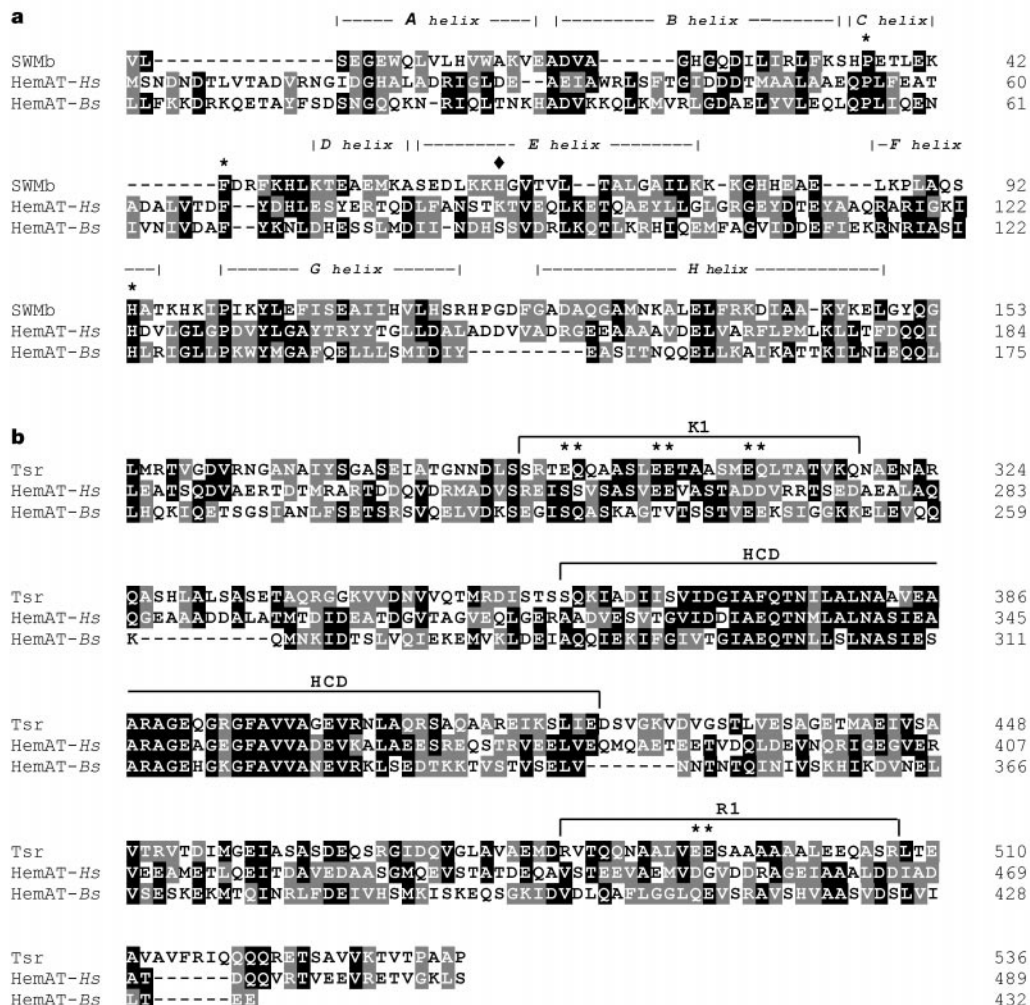


Figure 1 The conserved sequences within HemAT-*Hs*, HemAT-*Bs*, sperm-whale myoglobin (SWMb) and Tsr. Black boxes indicate positions at which the residues are identical, and grey boxes highlight residues that are similar. Sequences were aligned using the Clustal program of the MegAlign/DNASTAR. **a**, Alignment of the amino-terminal domain of HemAT-*Hs*, HemAT-*Bs* and SWMb. Helical regions in SWMb (helices A–H) are delineated by dotted lines. Pro (P), Phe (F) and His (H93) residues in SWMb are marked

with asterisks; Distal His (H64) in myoglobin is marked with a diamond. **b.** Alignment of the carboxyl-terminal domains of HemAT-*Hs*, HemAT-*Bs* and Tsr from *E. coli*. The extent of the two-methylation regions, K1 and R1, and the highly conserved domain (HCD) of Tsr are indicated. The Glx—Glx (EE or EQ) doublets corresponding to methylation sites of Tsr are marked with asterisks.

constructed *hemATs* deletion and overexpression strains. We analysed the behaviour of wild-type strains of both species, of single-deletion strains of *H. salinarum*¹⁴ (HemAT-*Hs* was once known as HtrX) and of the multiple-transducer-deleted Δten strains of *B. subtilis*. Aerotaxis assays were performed in flat microcapillaries, observed with a dark-field microscope coupled to a time-lapse digital video system (D.B. and M.A., unpublished data). Wild-type *H. salinarum* strain Flx15 forms two clear aerotactic bands; an aerophilic band near the air interface and an aerophobic band farther from the interface (Fig. 4a). Formation of the aerophilic band is mediated by the HtrVIII transducer¹⁵. If the aerophobic band forms in response to signals from HemAT-*Hs*, this band should not form with a $\Delta hemAT-Hs$ strain, as observed (Fig. 4a). Furthermore, the aerophobic response is clearly seen in the $\Delta htrVIII$ strain when HemAT-*Hs* is present (Fig. 4a). These data also indicate that the aerophobic response of wild-type cells is partially masked by the strong aerophilic response. We conclude that HemAT-*Hs* is essential for the aerophobic response in *H. salinarum*.

The aerophilic response in *B. subtilis* proceeds more rapidly than in *H. salinarum* (30 compared with 180 min) because *B. subtilis* swims faster than *H. salinarum*. In the wild-type, an aerotactic band formed at the air interface (Fig. 4b). This band did not form in a strain from which all 10 putative MCP-like transducers (Δten) were deleted (Fig. 4b). A strain lacking only HemAT-*Bs* showed an aerophobic response, indicating the presence of a second, unidentified aerotaxis receptor (data not shown). To demonstrate the physiological function of HemAT-*Bs* unequivocally, we overexpressed *hemAT-Bs* in a strain from which all *B. subtilis* transducer genes were deleted (Δten strain). When HemAT-*Bs* was overexpressed in the Δten strain, the aerophilic response was observed (Fig. 4b). These assays demonstrate that HemAT-*Bs* is involved in an aerophilic response in *B. subtilis*.

In contrast to *E. coli*, adaptation during aerotaxis in *H. salinarum* and *B. subtilis* is a methylation-dependent process^{15–19}. Compared with the bacterial chemotransducer Tsr, HemAT-*Hs* has two putative methylation sites in its C terminus (both are in the K1 region) and HemAT-*Bs* has one in each of the K1 and R1 regions (Fig. 1b). To determine if these putative methylation sites in HemATs can be methylated by the CheR methyltransferase, *H. salinarum* and *B. subtilis* cells were radiolabelled with methionine tributated on the

methyl group ([methyl-³H]methionine) after blocking protein synthesis. The radiolabelled cells were processed for fluorography and immunoblotting with a polyclonal antibody raised against the highly conserved region of methyl-accepting transducers⁸. A single radiolabelled band missing from the $\Delta hemAT-Hs$ strain (Fig. 2b, lane 1) is present in the overexpression strain (Fig. 2b, lane 2). This band is also recognized by the antibody, suggesting that HemAT-*Hs* is indeed a methyl-accepting transducer (Fig. 2b, lanes 1', 2'). In contrast, we were unable to detect any [methyl-³H]-labelling in HemAT-*Bs* (data not shown). In addition to the HemAT-*Hs* band in the overexpression strain (lanes 2, 2'), a prominent band at above 86K lights up strongly with the anti-transducer antibody and intensely with [methyl-³H] labelling in the deletion strain. The expression of this transducer (most likely the sensory rhodopsin I transducer HtrI (ref. 8) based on its migration on SDS-PAGE) is relatively low when recombinant HemAT-*Hs* is expressed. It is likely that the presence or absence of the aerotaxis transducer HemAT-*Hs* in *H. salinarum* is related to the expression level of HtrI. Together

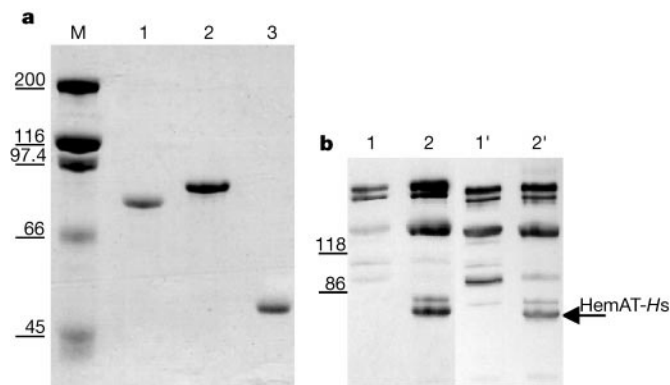


Figure 2 Characterization of HemAT proteins. **a**, Approximately 5 μ g of purified proteins was loaded in each lane during 10% SDS-PAGE²⁸. Lane 1, HemAT-*Hs*; lane 2, HemAT_{6xHis}-*Hs*; lane 3, HemAT-*Bs*. The M_r markers ($\times 10^3$) are shown on the left (lane M). **b**, Fluorograph and immunoblot of HemAT-*Hs*. Radiolabelling and immunoblotting were performed as previously described²⁸. Lane 1, fluorograph of proteins from the $\Delta hemAT-Hs$ strain; lane 2, fluorograph of proteins from the $\Delta hemAT-Hs/hemAT-Hs^{++}$ strain. Lane 1', immunoblot of $\Delta hemAT-Hs$ strain; and lane 2', immunoblot of $\Delta hemAT-Hs/hemAT-Hs^{++}$ strain using anti-transducer peptide antibody⁸. Bars indicate the positions of M_r markers ($\times 10^3$)

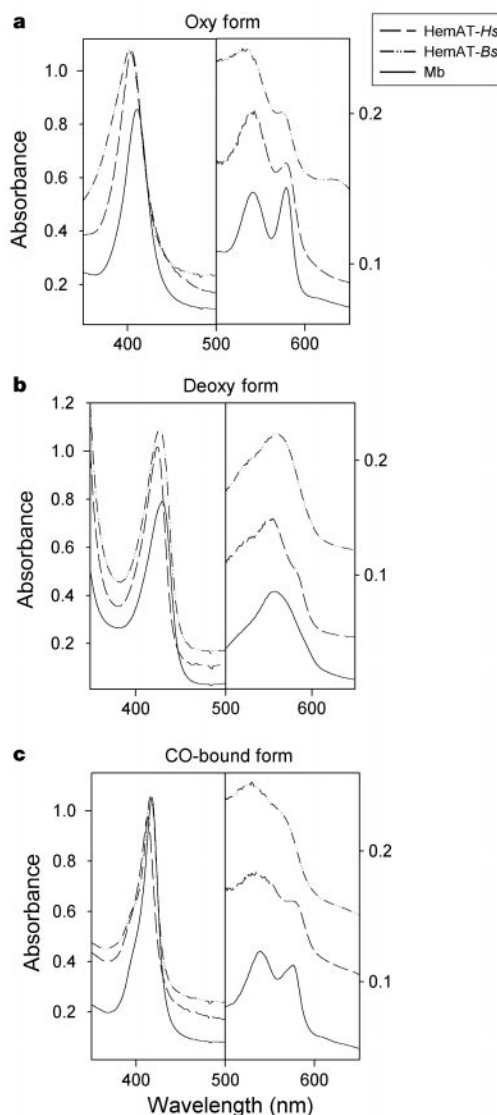


Figure 3 Absorption spectra of purified HemAT-*Hs*, HemAT-*Bs* and horse-heart myoglobin (HHMb). **a**, Oxygenated; **b**, deoxygenated and **c**, CO-bound forms of HemAT-*Hs*, HemAT-*Bs* and HHMb. The haem concentration in all samples was 20 μ M. Deoxygenated samples were prepared by adding sodium dithionite to the protein solutions. Spectra were obtained using a 1-cm quartz cuvette and a Milton-Roy Spectronic 3000 diode array UV/Vis spectrophotometer.

with the capillary assays, these data demonstrate an important difference in the signalling and adaptation mechanisms for aerotaxis mediated by HemAT-*Hs* and HemAT-*Bs*.

HemAT proteins constitute a new class of sensors that differ significantly from the known haem-containing O₂-sensor FixL (refs. 20, 21). FixL is a member of the large family of sensor kinases ubiquitous in bacterial two-component regulatory systems. Its haem-binding domain belongs to the PAS-domain superfamily^{18,19,22,23}. HemATs contain no PAS domains¹⁸ and differ from FixL both in spectral features and in physiological function. The absorption bands of HemATs are blue-shifted relative to FixL (415-nm Soret band), indicating that the proteins have distinct haem-pocket geometries. In addition, both HemATs participate in aerotaxis, whereas FixL regulates transcription. HemATs also differ from the aerotaxis transducer Aer in *E. coli*, which has a FAD-binding PAS domain^{24,25}.

We propose that the amino-terminal domains of HemATs act as sensors by binding diatomic oxygen at their haem when it is in the ferrous (Fe(II)) state. Oxygen binding presumably triggers a conformational change in the sensor domain that, in turn, alters the activity of the C-terminal signalling domain. The C-terminal domains of HemATs are very similar to the signalling domains of the MCP family of bacterial chemoreceptors, which associate with the cytoplasmic CheW and CheA proteins to mediate chemotaxis. □

Methods

Expression and purification of recombinant HemATs

PCR primers with sequences flanking the *hemAT-Hs* gene from *H. salinarum* strain Flx15 and encoding a *NdeI* or *BamHI* restriction site were used to amplify and clone the chromosomal gene into the pET expression vector (Novogen). The PCR product was initially ligated into the pCR-Blunt II TOPO cloning vector (Invitrogen) and then subcloned into pET-3b after digestion of the donor and recipient plasmids with *NdeI* and *BamHI*. PCR primers with sequences flanking the *hemAT-Bs* gene from *B. subtilis* strain OI1085 and encoding a *BamHI* or *PstI* restriction site were used to amplify and clone the chromosomal gene into the pCR-Blunt II TOPO vector. This fragment was later subcloned into the pMALcII expression vector (New England Biolabs). The resulting plasmids from both constructions were introduced into the *E. coli* BL21pLysS strain for protein expression.

BL21 pLysS host cells harbouring plasmids carrying the *hemAT-Hs* or *hemAT-Bs* genes were grown in Luria-Bertani broth with appropriate antibiotics, and synthesis of the proteins was induced with 0.6 mM isopropyl-β-D-thiogalactopyranoside. After a 2-h induction, the cells were collected by low-speed centrifugation (4,000g) at 4 °C for 15 min. The pellets were resuspended in buffer (50 mM NaCl, 50 mM Tris-HCl, pH 6.0) and sonicated for 4 min (12 pulses of 20 s with 30-s pauses). The cell lysate was centrifuged at

28,000g for 20 min. The red supernatant became the source of proteins for purification. The HemAT-*Hs* supernatant was applied to an anion-exchange POROS HQ/M column. A linear gradient of NaCl (0–1,500 mM) was applied, and HemAT-*Hs* eluted at about 400 mM. Fractions containing HemAT-*Hs* (monitored by the Soret-band absorbance at 410 nm and SDS-PAGE) were further purified by a HiLoad Superdex 200 gel-filtration column.

A saturated (NH₄)₂SO₄ solution was added to 30% saturation to the HemAT-*Bs* supernatant and centrifuged at 28,000g for 20 min. The optically clear, light-red supernatant was further fractionated by adding (NH₄)₂SO₄ to 36% saturation, and the precipitate was pelleted by centrifugation. The pellet was resuspended in buffer (200 mM NaCl, 50 mM Tris-HCl, pH 8.0) and further purified by a HiLoad Superdex 75 column.

Expression and purification of His-tagged HemAT-*Hs* from *H. salinarum*

A plasmid encoding C-terminal six-His-tagged HemAT-*Hs* was constructed by two-step PCR. In the first step, six His codons were fused to HemAT-*Hs* immediately in front of the natural stop codon. In the second step, an *XbaI* restriction site was introduced at the 3' end of the gene. The second PCR product was subcloned into the *NdeI* and *XbaI* sites of plasmid pKJ427. This plasmid was introduced into a Δ hemAT-*Hs* strain¹⁴. Cells grown at 37 °C to mid-logarithmic phase were collected by centrifugation (4,000g) at 4 °C. The pellet was resuspended in buffer (200 mM NaCl, 50 mM Tris-HCl, pH 8.0) and sonicated for 3 min (12 pulses of 15 s with 20-s pauses). The cell lysate was centrifuged (100,000g) at 4 °C for 30 min, and the supernatant was used for purification. The POROS MC/M affinity column was charged with 100 mM CoCl₂. HemAT_{6xHis}-*Hs* was eluted with a linear gradient of imidazole (0–250 mM) and further purified using a HiLoad Superdex 200 gel-filtration column.

Overexpression construction of HemATs in native hosts

NdeI and *XbaI* restriction sites were engineered respectively into forward and reverse primers for PCR amplification of the *hemAT-Hs* gene. The PCR product was initially cloned into the pCR-Blunt II TOPO cloning vector, and the resultant plasmid was digested with *NdeI* and *XbaI* restriction enzymes. The digested fragment carrying the *hemAT-Hs* gene was subcloned into the halobacterial expression vector pKJ427. The final recombinant plasmid was transformed into the Δ hemAT-*Hs* strain.

To construct the Δ ten strain lacking all transducers, strains lacking *yhfV*, *yfmS*, *yohA* or *yvaQ* (ref. 9) were constructed using PCR of an internal fragment, ligating it into pHV501 (ref. 26), and integrating it into the chromosome, with selection for erythromycin. Mutant strains lacking *mcpC*, *tlpA* or the combination of contiguous genes *mcpA*, *mcpB*, *tlpA* and *tlpB* were already available. The mutations were combined into a single strain by introducing a nearby nutritional marker by cotransformation with pEB112 (ref. 27), selection for kanamycin, subsequent segregation of the plasmid, and finally co-transduction of the respective inactivated receptor gene with selection for prototrophy of the nearby nutritional marker. In the case of *yvaQ* and the combination *mcpA*, *mcpB*, *tlpA* and *tlpB*, selection of a mutation in *thrC* was first carried out using an integrated plasmid, with selection for spectinomycin.

To construct an expression plasmid, *BamHI* and *PstI* restriction sites were used to clone the *hemAT-Bs*, including its promoter and ribosome-binding site, into pEB112. The PCR product was initially cloned into the pCR-Blunt II TOPO vector and later subcloned into plasmid pEB112. The resulting plasmid was introduced into the Δ ten strain.

Received 16 September; accepted 25 November 1999.

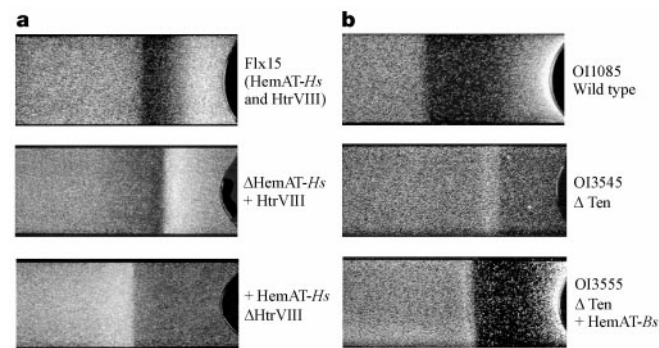


Figure 4 Aerotactic responses in *H. salinarum* and *B. subtilis*. **a**, *H. salinarum* strain Flx15 (HtrVIII and HemAT-*Hs* present) and mutant strains Δ hemAT-*Hs* (HtrVIII present) and Δ htrVIII (HemAT-*Hs* present). **b**, Wild-type *B. subtilis* strain OI1085 and mutant strains OI3545 (Δ ten) and OI3555 (overexpression of *hemAT-Bs* in Δ ten). Microcapillaries were filled halfway with mid-logarithmic-phase cell suspension, sealed at both ends and placed on a microscope stage pre-warmed to 35–37 °C. Time-lapse, dark-field microscopic images were recorded using a video-digitized camera linked to a computer. Bacteria appear as bright regions against a dark background. The images shown were taken at 180 min for *H. salinarum* and at 30 min for *B. subtilis*. The meniscus at the air interface is visible to the right in each image.

- Wakabayashi, S., Matsubara, H. & Webster, D. A. Primary sequence of a dimeric bacterial haemoglobin from *Vitreoscilla*. *Nature* **322**, 481–483 (1986).
- Potts, M., Angeloni, S. V., Ebel, R. E. & Bassam, D. Myoglobin in a Cyanobacterium. *Science* **256**, 1690–1691 (1992).
- Anderson, C. R., Jensen, E. O., Llewellyn, D. J., Dennis, E. S. & Peacock, W. J. A new hemoglobin gene from soybean: a role for hemoglobin in all plants. *Proc. Natl Acad. Sci. USA* **93**, 5682–5687 (1996).
- Hardison, R. Hemoglobins from bacteria to man: evolution of different patterns of gene expression. *J. Exp. Biol.* **201**, 1099–1117 (1998).
- Suzuki, T. & Imai, K. Evolution of myoglobin. *Cell Mol. Life Sci.* **54**, 979–1004 (1998).
- Hardison, R. The Evolution of Hemoglobin. *Am. Sci.* **87**, 126–137 (1999).
- Rodgers, K. R. Heme-based sensors in biological systems. *Curr. Opin. Chem. Biol.* **3**, 158–167 (1999).
- Zhang, W., Brooun, A., McCandless, J., Banda, P. & Alam, M. Signal transduction in the archaeon *Halobacterium salinarum* is processed through three subfamilies of 13 soluble and membrane-bound transducer proteins. *Proc. Natl Acad. Sci. USA* **93**, 4649–4654 (1996).
- Kunst, F. *et al.* The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature* **390**, 249–256 (1997).
- Boyd, A., Kendall, K. & Simon, M. I. Structure of the serine chemoreceptor in *Escherichia coli*. *Nature* **301**, 623–626 (1983).
- Bashford, D., Chothia, C. & Lesk, A. M. Determinants of a protein fold. Unique features of the globin amino acid sequences. *J. Mol. Biol.* **196**, 199–216 (1987).
- Vinogradov, S. N., Walz, D. A. & Pohajdak, B. Organization of non-vertebrate globin genes. *Comp. Biochem. Physiol.* **103**, 759–773 (1992).
- Ihara, K. & Mukohata, Y. The ATP synthase of *Halobacterium salinarum* (halobium) is an archaeobacterial type as revealed from the amino acid sequences of its two major subunits. *Arch. Biochem. Biophys.* **286**, 111–116 (1991).
- Brooun, A. *Primary Structures and Functional Analysis of the Four Transducers: HtrIV, HtrVIII, HtrXI and HtrXII from the Archaeon Halobacterium salinarum*. Thesis, Univ. Hawaii (1997).
- Brooun, A., Bell, J., Freitas, T., Larsen, R. W. & Alam, M. An archaeal aerotaxis transducer combines subunit I core structures of eukaryotic cytochrome c oxidase and eubacterial methyl-accepting chemotaxis proteins. *J. Bacteriol.* **180**, 1642–1646 (1998).
- Lindbeck, J. C., Goulbourne, E. A., Johnson, M. S. & Taylor, B. L. Aerotaxis in *Halobacterium salinarum* is methylation-dependent. *Microbiology* **141**, 2945–2953 (1995).

17. Wong, L. S., Johnson, M. S., Zhulin, I. B. & Taylor, B. L. Role of methylation in aerotaxis in *Bacillus subtilis*. *J. Bacteriol.* **177**, 3985–3991 (1995).
18. Taylor, B. L. & Zhulin, I. B. PAS domains: internal sensors of oxygen, redox potential, and light. *Microbiol. Mol. Biol. Rev.* **63**, 479–506 (1999).
19. Zhulin, I. B. & Taylor, B. L. Correlation of PAS domains with electron transport-associated proteins in completely sequenced microbial genomes. *Mol. Microbiol.* **29**, 1522–1523 (1998).
20. Gilles-Gonzalez, M. A., Ditta, G. S. & Helinski, D. R. A haemoprotein with kinase activity encoded by the oxygen sensor of *Rhizobium meliloti*. *Nature* **350**, 170–172 (1991).
21. Monson, E. K., Weinstein, M., Ditta, G. S. & Helinski, D. R. The FixL protein of *Rhizobium meliloti* can be separated into a heme-binding oxygen-sensing domain and a functional C-terminal kinase domain. *Proc. Natl Acad. Sci. USA* **89**, 4280–4284 (1992).
22. Zhulin, I. B., Taylor, B. L. & Dixon, R. PAS domain S-boxes in *Archaea*, *Bacteria* and sensors for oxygen and redox. *Trends Biochem. Sci.* **22**, 331–333 (1997).
23. Gong, W. *et al.* Structure of a biological oxygen sensor: a new mechanism for heme-driven signal transduction. *Proc. Natl Acad. Sci. USA* **95**, 15177–15182 (1998).
24. Rebbapragada, A. *et al.* The Aer protein and the serine chemoreceptor Tsr independently sense intracellular energy levels and transduce oxygen, redox, and energy signals for *Escherichia coli* behavior. *Proc. Natl Acad. Sci. USA* **94**, 10541–10546 (1997).
25. Bibikov, S. I., Biran, R., Rudd, K. E. & Parkinson, J. S. A signal transducer for aerotaxis in *Escherichia coli*. *J. Bacteriol.* **179**, 4075–4079 (1997).
26. Vagner, V., Dervyn, E. & Ehrlich, S. D. A vector for systematic gene inactivation in *Bacillus subtilis*. *Microbiology* **144**, 3097–3104 (1998).
27. Leonhardt, H. & Alonso, J. C. Construction of a shuttle vector for inducible gene expression in *Escherichia coli* and *Bacillus subtilis*. *J. Gen. Microbiol.* **134**, 605–609 (1988).
28. Alam, M. & Hazelbauer, G. L. Structural features of methyl-accepting taxis proteins conserved between archaeobacteria and eubacteria revealed by antigenic cross-reaction. *J. Bacteriol.* **173**, 5837–5842 (1991).

Acknowledgements

We thank H. G. deCouet, S. Donachie, G. L. Hazelbauer, M. Manson and O. R. Zaborsky for helpful comments on the manuscript, A. Brooun for the *hemAT-Hs* deletion strains, H. Chen and J. Yang for participation in the initial stage of protein purification, T. Freitas for preparing Fig. 1, and J. Spudich for the shuttle vector pKJ427. This investigation was supported by National Science Foundation CAREER grant to M.A. HemAT-Hs (HtB) has been given GenBank accession number U75436, and HemAT-Bs (YhfV) GenBank accession number Y14084.

Correspondence and requests for materials should be addressed to M.A. (e-mail alam@hawaii.edu).

Motor disorder in Huntington's disease begins as a dysfunction in error feedback control

Maurice A. Smith*, Jason Brandt† & Reza Shadmehr*

* Department of Biomedical Engineering and † Department of Psychiatry and Behavioral Sciences, Johns Hopkins University, Baltimore, Maryland 21205-2195, USA

A steady progression of motor dysfunction takes place in Huntington's disease¹ (HD). The origin of this disturbance with relation to the motor control process is not understood. Here we studied reaching movements in asymptomatic HD gene-carriers (AGCs) and subjects with manifest HD. We found that movement jerkiness, which characterizes the smoothness and efficiency of motion, was a sensitive indicator of presymptomatic HD progression. A large fraction of AGCs displayed elevated jerk even when more than seven years remained until predicted disease onset. Movement termination was disturbed much more than initiation and was highly variable from trial to trial. Analysis of this variability revealed that the sensitivity of end-movement jerk to subtle, self-generated early-movement errors was greater in HD subjects than in controls. Additionally, we found that HD corrective responses to externally-generated force pulses were greatly disturbed, indicating that HD subjects display aberrant responses to both external and self-generated errors. Because feedback corrections are driven by error and are delayed such that they

predominantly affect movement termination, these findings suggest that a dysfunction in error correction characterizes the motor control deficit in early HD. This dysfunction may be observed years before clinical disease onset and grows worse as the disease progresses.

Huntington's disease is an autosomal dominant inherited neurologic disorder caused by a glutamate repeat expansion in the IT15 gene². Disease symptoms appear in the fourth or fifth decade of life¹,

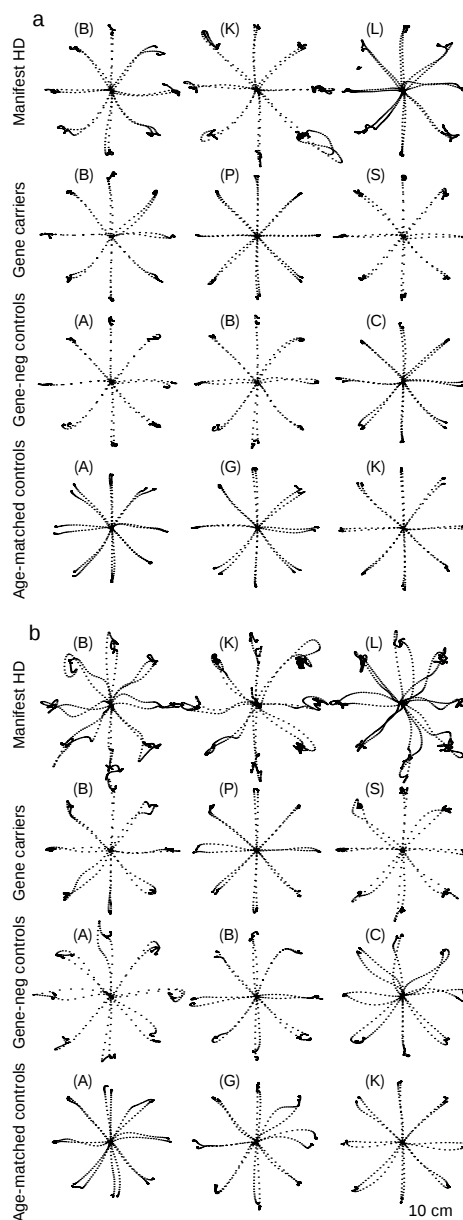


Figure 1 Hand paths from selected subjects after 200 practice trials (movements 201–300). **a**, The two most regular movements in each direction. **b**, The two least regular movements in each direction. Movement regularity was determined by the correlation coefficient⁶ of the velocity profile with the velocity profile of that subject's typical movement. The typical movement was defined as the movement in each direction with the highest average correlation to other movements. Hand paths are plotted from the centre out relative to their starting positions. Points are spaced 30 ms apart in time. The distance between consecutive points is proportional to the movement speed during that interval. Top row, subjects with manifest HD. Second row, asymptomatic gene-carriers. Third row, controls who have a parent with HD but who are mutation negative. Bottom row, controls age-matched to the asymptomatic gene-carriers. The letter that labels each subject identifies him or her within each group in Fig. 2.

and then progress steadily during the succeeding 10–20 years. Significant, progressive atrophy of the basal ganglia is evident in HD (ref. 3), and may even be detectable before clinical onset⁴. Its identifying motor sign is chorea: the occurrence of rapid, irregular, and arrhythmic complex involuntary movements. Along with chorea, substantial impairment of voluntary movement also occurs¹, and it may be of greater functional importance in the lives of patients⁵. Two control processes take part in the execution of arm movements. Feedforward control is the generation of motor commands based a priori on the desired action and an internal model of the system's response^{6,7}, whereas feedback control entails the 'mid-flight' correction of these commands based on errors detected during their execution⁸. To study these processes in HD, we used a high-performance manipulandum⁹ to record visually guided reaching arm movements in patients with HD and pre-symptomatic individuals with the HD mutation up to 18 years from predicted disease onset.

Figure 1 displays hand paths of the two most regular and two least regular movements in each direction, for several subjects after 200 practice movements. The range of movement quality is generally much larger in HD patients than in controls. Some movements made by symptomatic HD subjects appear normal whereas others are markedly irregular. Specifically, many HD movements have large changes in direction, smooth or abrupt, when approaching the target, whereas almost all movements fail to stop efficiently and smoothly. One way to characterize these irregularities is to quantify the smoothness of each movement. Smoothness can be defined as the lack of abrupt change; thus a trajectory that minimizes abrupt changes in a variable will maximize its smoothness. Human reaching movements are of near-maximal acceleration smoothness, quantified by minimal cumulative squared acceleration change¹⁰, also referred to as minimum total squared jerk.

Figure 2a shows that all HD patients and several AGCs made movements with above-normal jerk. All HD subjects reduced their movement jerk with practice, although not nearly to control levels. The smoothness of movements increased over time even in subjects who were markedly more jerky than normal. To compare the initiation and completion of movement, we split each movement at the peak in the speed profile. This is roughly the point at which movement towards the target switches from acceleration to deceleration. Comparison of the total squared jerk before and after the peak in the speed profile (Fig. 2b) reveals that although fewer than half of the HD patients have high jerk during both parts of the movement, all have above-normal post-peak jerk. Only 2 of 16 AGCs show above-normal pre-peak jerk, but most have high post-peak jerk. To assess disease progression in AGCs, we used an estimate of disease onset age based on each subject's parental onset age and glutamate repeat length⁴. The amount of post-peak jerk correlated significantly ($r = -0.62$) with estimated time to disease onset for AGCs whereas the pre-peak jerk did not ($r = 0.02$). Post-peak jerk was above normal in 4 of 5 close-to-onset (<7 years) subjects and in 3 of 9 far-from-onset subjects (>7 years). As groups, both HD patients and AGCs had significantly higher than normal post-peak jerk ($P < 10^{-6}$, $P < 0.00012$; see Fig. 2e). Moreover, AGCs who were close to predicted disease onset had significantly higher jerk in the third set than far-from-onset subjects ($P < 0.0062$) who in turn had significantly higher jerk than controls ($P < 0.013$).

Figure 2c shows that initial aiming is not greatly disturbed in HD. Average directional aiming bias (see Methods section) is normal in 9 of 11 HD patients. Aiming variability is in the control range for a majority (6 of 11) of patients, although, as a group, subjects with manifest HD display increased aiming variability. All AGCs have normal aiming biases of 3–7° and none has above-normal aiming variability.

These results suggest that HD movements often begin normally, but become jerky and irregular at some point during their course. Comparison of the average time course of raw squared jerk profiles

between groups for movements in different speed ranges (Fig. 3) reveals a strikingly consistent pattern. At each speed range, the HD jerk profile closely matches the control profile during the beginning of movement, but begins to separate from it 200–300 ms after onset. Note that the end-movement squared jerk is 10–30 times greater in HDs than controls.

Why do HD movements begin to become irregular 200–300 ms into their course and not before? Corrective actions based on visual^{11,12} and proprioceptive¹³ information acquired during reaching movements begin to take place at about the time at which HD movements become irregular, so one possibility is that the system that generates these corrective actions is disturbed. Alternatively, there may be a disturbance in decelerating movement or another end-movement process independent of error correction.

To distinguish between these possibilities, we characterized the performance of the error feedback control system in HD. If the system that generates corrective actions during movement (the feedback controller) is affected in HD, then a strong relationship should exist between magnitude of error early in the movement and size of disturbance later. Error, the difference between desired system state and actual state, is the primary input to a feedback controller. If the error during a movement is small, then large

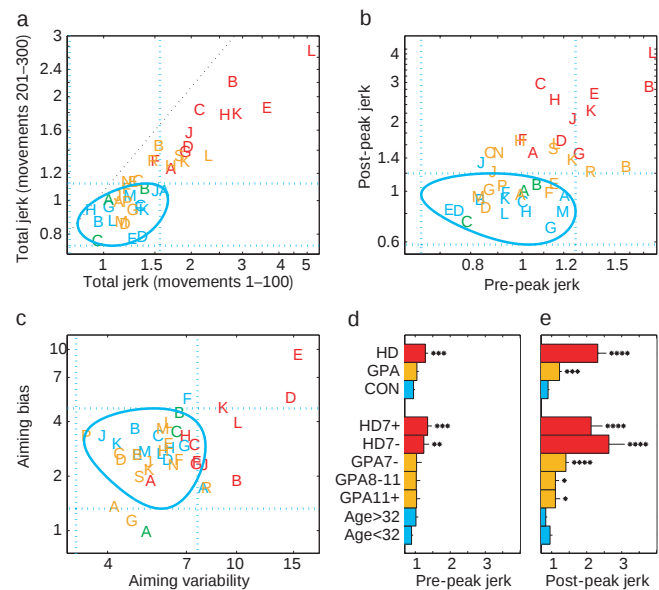


Figure 2 Quantification of movement properties. Red, subjects with manifest HD. Orange, asymptomatic gene-carriers. Blue, age-matched controls. Green, gene-negative individuals having a parent with HD. **a–c** are plotted on a logarithmic scale for clarity. The blue dotted lines show the 95% confidence intervals for the control distribution (mean ± 1.96 standard deviations). The black diagonal line in **a** represents the axis of equality ($y = x$). **a**, The mean of the normalized total squared jerk for all subjects in the first 100 movements and in movements 201–300. All symptomatic subjects and a subset of asymptomatic gene-carriers have higher than normal jerk. **b**, The mean of normalized jerk (in movements 201–300) before and after the peak in the movement speed. The pre-peak segment, which reflects movement initiation, appears much less disturbed in HD than the post-peak segment, which reflects movement completion. **c**, Aiming bias and aiming variability. We refer to aiming as the direction of travel with respect to the initial target direction, during the pre-peak movement segment. Like pre-peak jerk, aiming reflects the quality of movement. **d,e**, Group-wise comparisons of normalized pre-peak and post-peak jerk, respectively. Asterisks indicate significantly worse performance than control subjects. * $P < 0.05$. ** $P < 0.01$. *** $P < 0.001$. **** $P < 0.0001$. HD7+, subjects with manifest HD for more than 7 years. HD7-, subjects with manifest HD for less than 7 years. GPA7-, gene-carriers less than 7 years from predicted onset. GPA7-11, gene-carriers 7 to 11 years from predicted onset. GPA11+, gene-carriers more than 11 years from predicted onset. Age > 32, controls more than 32 years old. Age < 32, controls less than 32 years old.

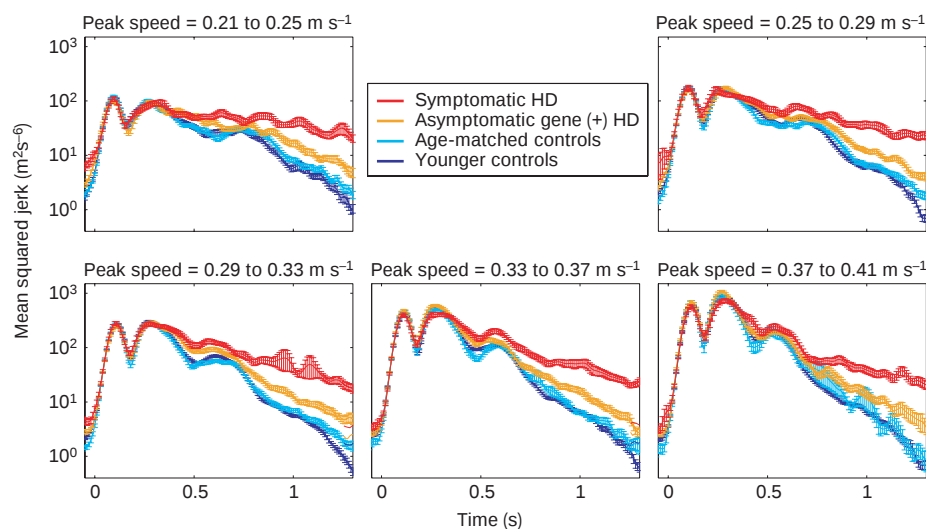


Figure 3 Squared jerk profiles for different movement speeds. Average raw squared jerk $\pm$ standard error is plotted on a logarithmic scale as function of time since movement onset for movements in each speed range, in different subject groups. For reference, the

second peak in the squared jerk profile approximately corresponds to the peak in the speed profile. Note that at all speed ranges the separate groups have quite similar jerk profiles until 300 ms, but HD subjects have very different profiles after this point.

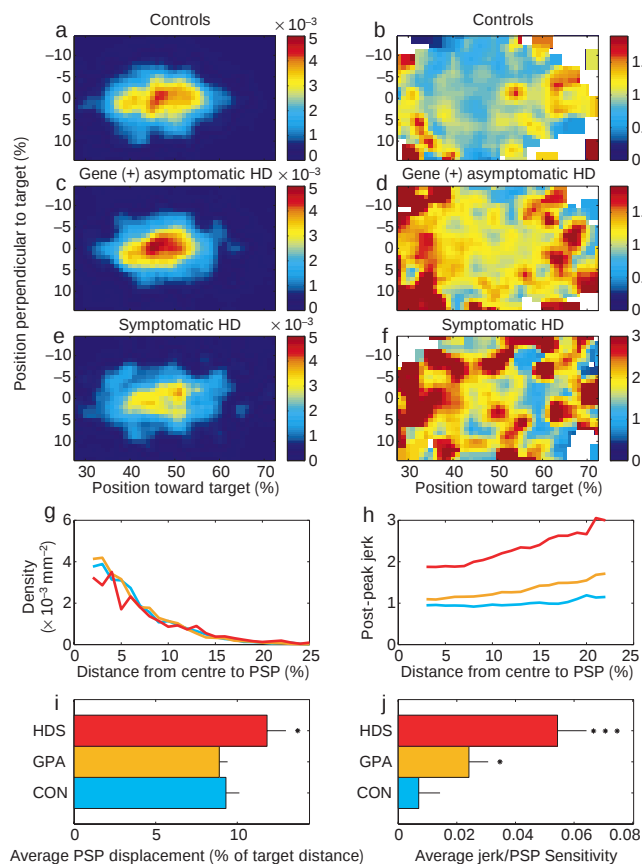


Figure 4 Errors that occur early in the movement, before the hand reaches its peak speed, predict jerk that occurs later. **a,c,e**, The two-dimensional probability densities of peak speed positions (PSPs) for each group during movements 301–400. This is the likelihood of the peak speed occurring at each position. The x-axes are in the target direction and the y-axes are in the direction perpendicular to the target. PSPs are clustered near the movement midway point. **g**, Summary of **a,c,e**: the average probability density at a given displacement between the PSP and movement midway point (50%, 0%). Large values for the distance from midway point to peak-speed position indicate large early-movement errors, which are uncommon. The similarity of these distributions indicates that there are not large differences in the pattern of error recorded from the three subject groups early in

the movement. **b,d,f**, End-movement jerk as a function of peak-speed position. When large early movement errors occur, as indicated by large PSP displacements (near the image boundaries), the post-peak jerk is increased for all groups, but the increase is greater for asymptomatic and symptomatic HD subjects than controls. White colour on images indicates no data. **h**, Summary of **b,d,f**: the average end-movement jerk at a given displacement between the PSP and movement midway point. **i**, Average PSP displacement for each group. **j**, Sensitivity of end-movement jerk to PSP, measured as the slope of the relationship between them. Asterisks indicate significantly worse performance than control subjects. * $P < 0.05$. ** $P < 0.01$. *** $P < 0.001$.

corrective actions are not necessary and dysfunction in feedback control might have only minimal effects. However, if errors are large, substantial corrective actions are required, and dysfunctional feedback control may cause significant problems with movement.

Although we can measure actual movements quite accurately, we have no reliable way of estimating the desired motion state within a single movement. This makes direct estimation of error infeasible. If errors are symmetrically distributed, then the average movement profile should approximate the average motion plan (that is, the average intended movement of the subject). However, subjects may plan to move slightly faster on one trial, then slower on the next, and these variations in the plan may be as large as the typical errors between plan and action. Although the motor plan might change across trials, a few of its properties are invariant. Velocity profiles of reaching movements are symmetric and unimodal: the peak in the velocity profile occurs very close to the movement's midway point, regardless of amplitude, direction or speed¹⁴. Therefore, the distance between the midway point and the position where the peak speed occurs can be used as an indicator of error early in the movement.

The relationship between error early in movement and jerk late in movement is shown in Fig. 4. The probability distributions of peak-speed positions (PSPs) do not differ markedly between groups (Fig. 4 a,c,e,g and i). This implies that amount of error early in the movement is comparable between groups. However, the reaction to this error is very different between groups. Figure 4 b,d,f,h and j shows that the jerk that occurs after the hand reaches the PSP varies considerably with PSP in all subject groups. When the PSP is far from the midway point, indicating large error early in the movement, post-peak jerk is high. However, the sensitivity of post-peak jerk to PSP is much greater in HD patients and AGCs than controls. Figure 4h shows that the difference in mean post-peak jerk between

controls and AGCs is much greater at large PSP displacements than at small displacements. Subjects with manifest HD ($P = 0.006$) and AGCs ($P = 0.04$) display significantly higher sensitivity in their response of end-movement jerk to early-movement error than do controls, as measured by slope in this relationship. The sensitivity of post-peak jerk to PSP displacement indicates the degree to which end-movement smoothness depends on early-movement error. The increase in this sensitivity in HD subjects with and without clinically manifest symptoms suggests that an error-dependent control process is disturbed early in the disease course and further deteriorates with disease progression.

Previous reports have shown that cortical sensorimotor pathways are affected in patients with manifest HD. Whereas short-loop reflexes that are mediated by spinal mechanisms are normal, long-loop reflexes, which involve the transfer of proprioceptive information through cortical pathways, are reduced or absent in HD^{15,16}. Cortical responses to peripheral nerve stimulation, as measured by somatosensory evoked potentials (SEPs), are also reduced in HD^{15,17,18}, suggesting that diminished cortical sensory input is responsible for the long-loop reflex reduction¹⁶. In addition, a strong correlation exists between SEP deterioration and striatal glucose metabolism early in the course of HD¹⁷, hinting at a possible involvement of the striatum in the pathology of the SEPs and long-loop reflexes. As cortical sensorimotor pathways have a large role in mediating error correction during voluntary movement, disruption of these pathways could lead to dysfunctional feedback control during movement.

Our analysis of the response to internal, self-generated errors suggested that error correction in general might be disturbed in HD. To test this hypothesis we externally imposed errors upon movements of these subjects via an occasional brief (70 ms) force pulse shortly after movement initiation. These pulses were given

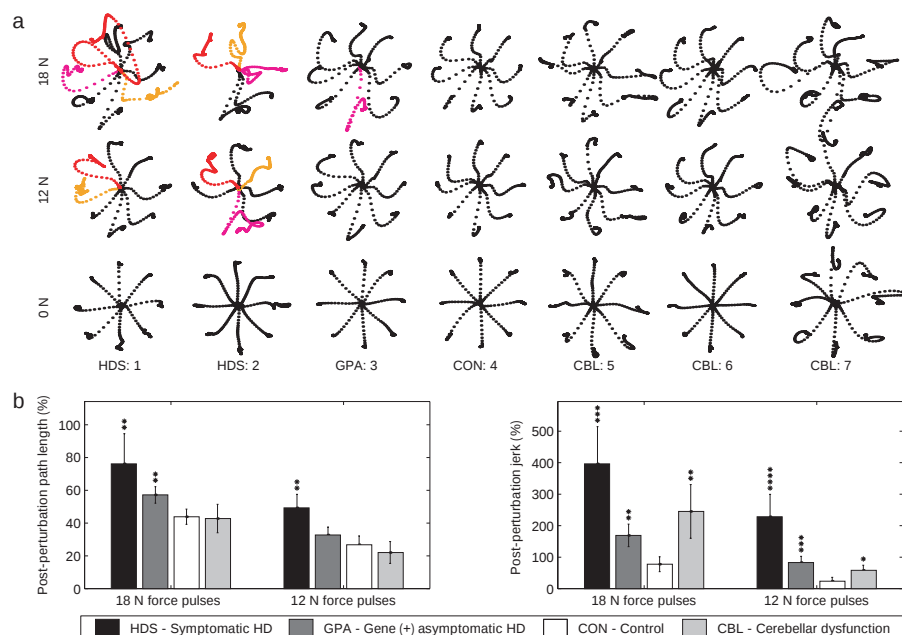


Figure 5 Force-pulse perturbations disturb the movements of HD subjects more than controls or cerebellar subjects. **a**, Sample movement trajectories during the perturbation task. Top two rows, movements during which an 18 or 12 N force pulse was applied in the 7:30 direction. Bottom row, typical unperturbed movements interspersed within the perturbed movements. Some of the movements made by HD subjects become greatly disturbed after a perturbation is applied. A few of the most greatly disturbed trajectories are highlighted. Note the trajectory towards 10:30 for HD subject 1. Here the movement begins in the correct direction but is perturbed to the left. The 'correction' then brings the

movement quickly back all the way past the original starting location, turns around and heads roughly towards the target, then makes a final turn, moves towards the target and stops. **b**, Change in movement properties in response to force-pulse perturbations (% change with respect to unperturbed movements). Right panel, normalized post-perturbation path length. Left panel, normalized post-perturbation jerk. Asterisks indicate significant increases above control disturbance. * $P < 0.05$. ** $P < 0.025$. *** $P < 0.01$. **** $P < 0.001$.

randomly on a minority of trials, and could be in eight different directions (indicated by the times corresponding to clockface directions) and of three different magnitudes (6, 12 and 18 N peak force). In addition to subjects with HD, we studied the performance of individuals with cerebellar deficits on this task. Hand paths of two symptomatic HD, one AGC, one control and three cerebellar subjects for movements perturbed with force pulses of 12 and 18 N are shown in Fig. 5a. These perturbations are substantial: trajectories can be altered by up to 5 cm (movement distance is 10 cm). Control subjects correct the perturbations smoothly and efficiently, but extremely large successive overshoots are seen during certain HD movements. Not all HD movement corrections appear irregular, but some are so markedly disturbed that they bear little qualitative similarity to any movements that we recorded from controls.

Several movements made by cerebellar subjects appear to have irregular and inefficient corrections, but close inspection of many of these trajectories reveals a striking similarity between the pattern of overshoots in these movements and the overshoots present in unperturbed movements in the same direction. This suggests that the irregularity that exists in these perturbed movements has a stronger relationship to the direction of movement than to the presence of perturbation. We note that such movements toward 3:00, 4:30 and 7:30 for subjects 5 and 7.

HD subjects appear to have dysfunctional reactions to movement errors caused by external perturbations. During the post-perturbation movement segment (the corrective period), jerk and path length relatively increase over unperturbed movements significantly more for HD subjects than for controls (see Fig. 5b) for both 12- and 18-N perturbations. The error-correction performance of AGCs generally falls between that of controls and subjects with manifest HD. Cerebellar subjects, like symptomatic HD patients, had worse unperturbed movement performance than controls, but the decrement in their performance when perturbations were given was generally more like that of controls than HD subjects. This suggests that subjects with HD generally have greater deficits in error feedback control than do cerebellar patients.

Previously, assessments of motor, cognitive or psychiatric function in HD have revealed only subtle deficits in presymptomatic subject groups^{19–21} if any^{22,23}. When changes have been detected, they were not sufficiently specific to permit discrimination between mutation-positive and mutation-negative individuals, or even reliable identification of people with early-stage manifest HD. In contrast, low glucose metabolism in the basal ganglia has been reported in about two-thirds of asymptomatic at-risk individuals tested^{17,24,25} and basal ganglia volumes may be reduced years before clinical onset⁴, suggesting that brain pathology precedes manifestation of the behavioural dysfunction previously studied. However, end-movement jerk appears to be a sensitive indicator of HD progression, suggesting the existence of a direct behavioural correlate to the early brain pathology of HD.

Our results, which suggest relatively unaffected feedforward control but dysfunctional feedback control in HD, may help to explain the pattern of motor learning deficits reported: learning rotary pursuit, which involves long continuous movements under closed-loop feedback control, is more impaired than mirror tracing with short discrete movements that are more open-loop in nature²⁶. Real-time error feedback control presents a formidable challenge to the CNS because of the large sensorimotor delays that occur²⁷. One way to improve the performance of such a time-delayed system is to include a component that can effectively predict away the delay and allow efficient error correction. This sort of predictor has been referred to as a forward model of system dynamics^{8,28}. Neurons in the basal ganglia have been shown to predict reward by firing vigorously in advance of reward upon completion of the requirements for reward attainment²⁹, hinting that predictive capacity may be a general feature of some basal ganglia structures. The main

output of the basal ganglia modulates the action of the thalamus, which relays sensory information to the cortex. This information stream is likely to participate in error feedback control. □

Methods

Subjects

Eleven patients positive for the IT-15 mutation and symptomatic with Huntington's disease, sixteen mutation-positive presymptomatic subjects, three mutation-negative subjects who had a parent with HD and twelve other age-matched controls participated in the first experiment. Five symptomatic and nine presymptomatic HD subjects, eight age-matched controls and six subjects with cerebellar lesions participated in the second experiment. All subjects used their dominant hand, and all but one presymptomatic subject in the first experiment were right-handed. The direct gene test for IT-15 mutation was conducted at the Johns Hopkins Huntington's Disease Project. The number of CAG (glutamate) trinucleotide repeats was determined, and subjects with >37 repeats were called mutation-positive. Subjects with <34 were called mutation-negative. All cerebellar subjects that we studied had been diagnosed clinically with cerebellar dysfunction, and all had lesions localized to the cerebellum on MRI. Four patients had generalized cerebellar atrophy, and two had suffered strokes of the right posterior inferior cerebellar artery (PICA). One of these patients also had a left PICA and a right superior cerebellar artery stroke.

Task

Subjects made quick reaching movements to targets spaced 10 cm away while grasping a lightweight two-joint manipulandum. The 1-cm square targets and a small cursor indicating the subject's hand position were displayed on a computer monitor in front of the subject⁶. We define the end of the movement as the beginning of the first time interval after movement onset when hand velocity always remained below a threshold of 0.03 m s⁻¹ for 200 ms. During the first experiment, the training took place in two sessions and within sessions it was subdivided into 100-movement sets. In the first session, subjects completed three sets of training. After a break of 3–4 hours, they began the second session, which consisted of a single set. The robot arm remained passive for all movements in both sets. During the second experiment the robot produced a 70-ms bell-shaped force pulse on a minority of randomly pre-selected trials (with a probability of one in four). The force pulse could be in any one of eight directions and of magnitude 6, 12 or 18 N. One force pulse in each direction and of each magnitude was given for each direction of movement. During both experiments we used a sling suspended from the ceiling to support the subject's upper arm in the horizontal plane. This helped to regularize the subjects' arm position and minimize the effort required to support the arm against gravity.

Analysis

We characterized aiming direction by defining angular aiming error as the difference between target direction and the direction of travel up to the peak speed point. Aiming error for a given movement can be positive or negative. Aiming bias for each subject was defined as the average of the magnitude of the average aiming error in each direction. Similarly, aiming variability for each subject was defined as the average of the standard deviation of aiming error in each direction.

Jerk is defined as the rate of change of acceleration with respect to time. To minimize the effect of discretization noise on the differentiation of the velocity signal, jerk was estimated by applying a Savitsky–Golay filter. The minimum total squared jerk (TSJ) required for a point-to-point movement is proportional to the fifth power of the movement speed and inversely proportional to the third power of movement excursion. Because of the strength of the relationship between these variables and TSJ, it is critical to normalize movement speed and excursion appropriately when comparing the amount of jerk between two movements. To account for the effect of speed and excursion on a movement's typical jerk, we normalized the TSJ measured for each movement by the average TSJ for movements made after practice by a large separate group of controls ($n=35$), at the appropriate speed and excursion. Similarly, we normalized the TSJ in the pre-peak and post-peak movement segments by the average values of these quantities for control movements of the same peak speed and excursion, and we refer to these quantities as pre-peak jerk and post-peak jerk.

As the trajectories of perturbed movements were often quite irregular, and the peak speed was often strongly influenced by the presence and direction of the perturbing force pulse, normalization of jerk by peak speed was no longer appropriate for these movements. Instead, we normalized the jerk measured after a perturbation offset for a given movement by dividing by the minimum possible jerk required to make the movement state transition characteristic of that movement segment. We refer to this quantity as post-perturbation jerk. Similarly, we define post-perturbation path length as the path length between the point of perturbation offset and the end point of movement divided by the straight-line distance between these two points. We compare the values of these quantities for movements during which perturbations were given to unperturbed movements. For unperturbed movements, we define the time point of 'perturbation offset' as that time when perturbation offset would have occurred had a perturbation been given.

Analytical methods

The position and velocity of the joints of the robot arm were recorded at 100 Hz from absolute and relative joint position encoders with resolution of 5.5×10^{-3} degrees and 8.0×10^{-4} degrees, respectively. This produced estimates of hand position and

velocity in cartesian coordinates with accuracy greater than 0.1 mm and 1.3 mm s⁻¹, respectively.

Jerk was estimated by applying a fourth-order Savitsky–Golay filter on a 250-ms window of velocity data. This filter is equivalent to taking the second derivative at the window's centre of the continuous least-squares best-fit fourth-order polynomial. This fourth-order polynomial fit is a low-pass filter with a cutoff frequency of 6.83 Hz. Power spectra of mean subtracted velocity profiles of very fast 10-cm reaching movements show that 99.9% of the power is below 6 Hz.

We assessed motion state transition efficiency using cumulative squared jerk to characterize the efficiency of recovery after perturbation offset. To accomplish this, we compared the amount of jerk that occurred between two different motion states within the same movement, with the jerk that would occur for a maximally smooth transition between those two states in the elapsed time. The minimum jerk trajectory between two motion states (state = [position, velocity, acceleration]) is given by a fifth-order polynomial in time:

$$x(t) = C_5(t/t_f)^5 + C_4(t/t_f)^4 + C_3(t/t_f)^3 + C_2(t/t_f)^2 + C_1(t/t_f) + C_0$$

Where position is represented by $x(t)$, t is time and t_f is the final time. C_k s are parameters that depend on the boundary motion states and on the time between them, t_f . They can be found by solving the boundary conditions on the motion state.

Once the coefficients are determined the cumulative squared jerk can be computed by simply integrating the squared jerk profile.

$$j(t) = x(t) = [60C_5(t/t_f)^2 + 24C_4(t/t_f) + 6C_3]/t_f^3$$

$$\text{Cumulative Squared Jerk} = \int_0^{t_f} j^2(t) dt$$

Received 26 August; accepted 3 November 1999.

1. Folstein, S. E. *Huntington's Disease a Disorder of Families* (Johns Hopkins Univ. Press, Baltimore, 1989).
2. The Huntington's Disease Collaborative Research Group. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**, 971–983 (1993).
3. Aylward, E. H. *et al.* Longitudinal change in basal ganglia volume in patients with Huntington's disease. *Neurology* **48**, 394–399 (1997).
4. Aylward, E. H. *et al.* Basal ganglia volume and proximity to onset in presymptomatic Huntington disease. *Arch Neurol* **53**, 1293–1296 (1996).
5. Brandt, J. *et al.* Clinical correlates of dementia and disability in Huntington's disease. *J. Clin. Neuropsychol.* **6**, 401–412 (1984).
6. Shadmehr, R. & Mussa-Ivaldi, F. A. Adaptive representation of dynamics during learning of a motor task. *J. Neurosci.* **14**, 3208–3224 (1994).
7. Shidara, M., Kawano, K., Gomi, H. & Kawato, M. Inverse-dynamics model eye movement control by Purkinje cells in the cerebellum. *Nature* **365**, 50–52 (1993).
8. Bhushan, N. & Shadmehr, R. Computational nature of human adaptive control during learning of reaching movements in force fields. *Biol. Cybern.* **81**, 39–60 (1999).
9. Shadmehr, R. & Brashers-Krug, T. Functional stages in the formation of human long-term motor memory. *J. Neurosci.* **17**, 409–419 (1997).
10. Flash, T. & Hogan, N. The coordination of arm movements: an experimentally confirmed mathematical model. *J. Neurosci.* **5**, 1688–1703 (1985).
11. Miall, R. C., Weir, D. J. & Stein, J. F. Manual tracking of visual targets by trained monkeys. *Behav. Brain Res.* **20**, 185–201 (1986).
12. Wolpert, D. M., Ghahramani, Z. & Jordan, M. I. Are arm trajectories planned in kinematic or dynamic coordinates? An adaptation study. *Exp. Brain Res.* **103**, 460–470 (1995).
13. Cordo, P. J. Kinesthetic control of a multijoint movement sequence. *J. Neurophysiol.* **63**, 161–172 (1990).
14. Atkeson, C. G. & Hollerbach, J. M. Kinematic features of unrestrained vertical arm movements. *J. Neurosci.* **59**, 2318–2330 (1985).
15. Meyer, B. U. *et al.* Motor responses evoked by magnetic brain stimulation in Huntington's Disease. *Electroencephalogr. Clin. Neurophysiol.* **85**, 197–208 (1992).
16. Noth, J., Podoll, K. & Friedemann, H. H. Long-loop reflexes in small hand muscles studied in normal subjects and in patients with Huntington's Disease. *Brain* **108**, 65–80 (1985).
17. Kuwert, T. *et al.* Comparison of somatosensory evoked potentials with striatal glucose consumption measured by positron emission tomography in the early diagnosis of Huntington's Disease. *Mov. Disord.* **8**, 98–106 (1993).
18. Topper, R., Schwarz, M., Podoll, K., Domges, F. & Noth, J. Absence of frontal somatosensory evoked potentials in Huntington's Disease. *Brain* **116**, 87–101 (1993).
19. Foroud, T. *et al.* Cognitive scores in carriers of Huntington's disease gene compared to noncarriers. *Ann. Neurol.* **37**, 657–664 (1995).
20. de Boo, G. M. *et al.* Early cognitive and motor symptoms in identified carriers of the gene for Huntington disease. *Arch. Neurol.* **54**, 1353–1357 (1997).
21. Siemers, E. *et al.* Motor changes in presymptomatic Huntington disease gene carriers. *Arch. Neurol.* **53**, 487–492 (1996).
22. Rothlind, J. C., Brandt, J., Zee, D., Codori, A. M. & Folstein, S. Unimpaired verbal memory and oculomotor control in asymptomatic adults with the genetic marker for Huntington's disease. *Arch. Neurol.* **50**, 799–802 (1993).
23. Blackmore, L., Simpson, S. A. & Crawford, J. R. Cognitive performance in UK sample of presymptomatic people carrying the gene for Huntington's disease. *J. Med. Genet.* **32**, 358–362 (1995).
24. Kuwert, T. *et al.* Striatal glucose consumption in chorea-free subjects at risk of Huntington's disease. *J. Neurol.* **241**, 31–36 (1993).
25. Mazzio, J. C. *et al.* Reduced cerebral glucose metabolism in asymptomatic subjects at risk for Huntington's Disease. *N. Engl. J. Med.* **316**, 357–362 (1987).

26. Gabrieli, J. D., Stebbins, G. T., Singh, J., Willingham, D. B. & Goetz, C. G. Intact mirror-tracing and impaired rotary-pursuit skill learning in patients with Huntington's Disease: Evidence for dissociable memory systems in skill learning. *Neuropsychology* **11**, 272–281 (1997).
27. Hogan, N. In *Multiple Muscle Systems* (eds Winters, J. M. & Woo, S. L. Y.) 149–164 (Springer, New York, 1990).
28. Miall, R. C. & Wolpert, D. M. Forward models for physiologic motor control. *Neur. Net.* **9**, 1265–1279 (1996).
29. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).

Acknowledgements

We thank K. Thoroughman and N. Bhushan. R.S. and M.S. conceived of the experiments, M.S. collected the data, and designed and carried out the data analysis guided by interaction with R.S., J.B., N.B. and K.T. M.S. and R.S. wrote the manuscript. The Johns Hopkins HD Center (director C. A. Ross) arranged patient visits and clinical assessment of HD patients. The work was supported by grants from the Whitaker Foundation (R.S.) and the National Institutes of Health (R.S. and J.B.), and a pre-doctoral fellowship from the National Institute of General Medical Sciences to M.S.

Correspondence and requests for materials should be addressed to M.S. (e-mail: msmith@bme.jhu.edu).

A neuronal analogue of state-dependent learning

D. E. Shulz⁺, R. Sosnik[†], V. Ego⁺, S. Haidarliu[†] & E. Ahissar[†]

[†] Department of Neurobiology, The Weizmann Institute of Science, 76100 Rehovot, Israel

⁺ Equipe Cognosciences, UNIC, Institut A. Fessard, CNRS, 91198 Gif sur Yvette, France

State-dependent learning is a phenomenon in which the retrieval of newly acquired information is possible only if the subject is in the same sensory context and physiological state as during the encoding phase¹. In spite of extensive behavioural and pharmacological characterization², no cellular counterpart of this phenomenon has been reported. Here we describe a neuronal analogue of state-dependent learning in which cortical neurons show an acetylcholine-dependent expression of an acetylcholine-induced functional plasticity. This was demonstrated on neurons of rat somatosensory 'barrel' cortex, whose tunings to the temporal frequency of whisker deflections were modified by cellular conditioning. Pairing whisker stimulation with acetylcholine applied iontophoretically yielded selective lasting modification of responses, the expression of which depended on the presence of exogenous acetylcholine. Administration of acetylcholine during testing revealed frequency-specific changes in response that were not expressed when tested without acetylcholine or when the muscarinic antagonist, atropine, was applied concomitantly. Our results suggest that both acquisition and recall can be controlled by the cortical release of acetylcholine.

The ascending cholinergic system³ has long been considered to be a candidate for mediating behavioural control of neuronal plasticity^{4–9}. This hypothesis is supported by behavioural and neurophysiological studies in the auditory^{10–13} and somatosensory systems^{14–18}. Whereas these studies demonstrated the permissive role of acetylcholine (ACh) during the induction of cortical plasticity^{10–14}, they did not address the possibility that ACh is also involved in the expression of the induced modifications. To examine this potential role of ACh, single- ($n = 99$) and multi-unit ($n = 85$) activities were recorded extracellularly from the barrel field¹⁹ of anaesthetized adult rats, using a multi-electrode array composed of one or two tungsten-in-glass electrodes and one combined electrode for recording and iontophoresis of ACh. Temporal-frequency tuning curves (TFTCs)

were obtained by mechanically deflecting the principal vibrissa at four different frequencies between 2 and 11 Hz (in a few cases 14 Hz was also applied), covering the frequency range predominantly used by the animal while exploring its environment²⁰. Typically, TFTCs of barrel cortex neurons show decreased spike counts and increased latencies with increasing frequencies²¹. The TFTC was determined again during ACh iontophoresis, and then pairing occurred. Pairing consisted of repetitive whisker deflection at one fixed frequency (5, 8 or 11 Hz) accompanied by ACh iontophoresis. After pairing, the TFTC was determined, once without ACh and once during application of ACh, thereby restoring the physiological conditions under which the pairing was carried out.

Pairing caused frequency-specific modification of the TFTCs that was expressed exclusively under ACh application. Three examples of significant increments in response, induced by pairing and revealed with ACh, are depicted in Fig. 1 (Fig. 1a–c, two-tailed Kolmogorov–Smirnov, $P < 0.0005$ compared with control). The principal whisker for each cell was stimulated at 5 (Fig. 1a), 8 (Fig. 1b) or

11 Hz (Fig. 1c) during pairing. The potentiation of the response was maximal for the conditioned frequency in each case and affected both the phasic (due to each whisker deflection) and tonic (due to the entire train of deflections) components of the response. The summed response (phasic + tonic per stimulus cycle) is referred to herein as ‘responsiveness’. No change was expressed when the responses were recorded without ACh (Fig. 1a–c, Kolmogorov–Smirnov test, $P > 0.1$). The modifications could be reversed by a second pairing using a different stimulation frequency. A new frequency-specific enhancement in response after a second pairing is shown in Fig. 1d. Usually, the second pairing resulted also in a significant reduction of response to the initially paired frequency (11 Hz, Fig. 1d) (see Supplementary Information). When the extinction of the effect was analysed by repeatedly testing the TFTC without and with ACh, the response modification was still statistically significant at least 45 min after the pairing but only under ACh (four out of four cells, two-tailed Kolmogorov–Smirnov, $P < 0.0005$). The consistent lack of expression without ACh, during periods that were interleaved with successful expression with ACh, excludes the possibility that the observed effects correspond to a delayed expression of cholinergic-induced plasticity²² (see Supplementary Information).

Overall, 33% (39 out of 119) of the single- and multi-units recorded with the combined electrode showed a statistically significant TFTC modification when tested with ACh after pairing. In contrast, when measured without ACh, TFTC changes were observed in fewer cases (21%, 25 out of 119; χ^2 test, $P < 0.05$). This difference was also valid for single units: 30% showed a modified TFTC (17 out of 57) when tested with ACh, whereas only 14% (8 out of 57) were modified when measured without ACh (χ^2 test, $P < 0.05$). Most of the changes expressed with ACh were response potentiations specific to the paired frequency (76%). Consistent with studies in the auditory cortex¹⁰, most of the effects observed during testing without ACh were decreases in responsive-

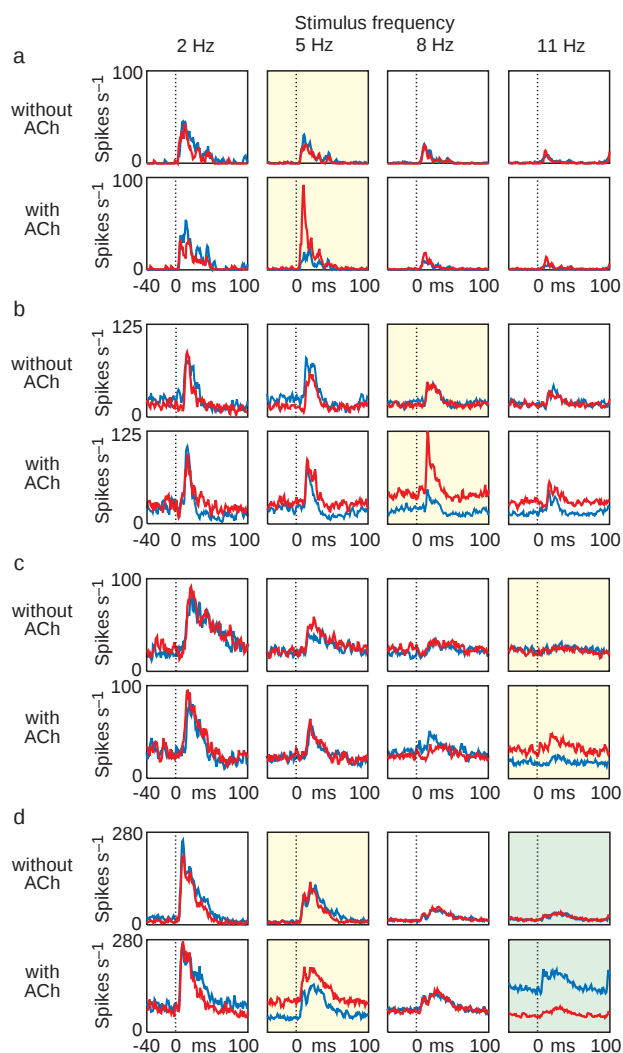


Figure 1 Plasticity of cortical responses expressed during ACh application in four different units. Peri-stimulus time histograms of responses before (blue) and after (red) pairing are superimposed. The activity preceding the stimulus onset (dashed lines) corresponds to the cell's tonic activation during the stimulus train. Yellow shading indicates the paired frequency. Pairing at 5 (a), 8 (b) and 11 Hz (c) resulted in an enhanced response to the paired frequency when tested with ACh (Kolmogorov–Smirnov, $P < 0.0005$), but not when tested without ACh ($P > 0.1$) (a–c). d, Reversal of the potentiation induced by a first conditioning at 11 Hz (green shading, $P < 0.0005$), by a second pairing at 5 Hz (yellow shading, $P < 0.0005$). No change was revealed when testing without ACh ($P > 0.9$).

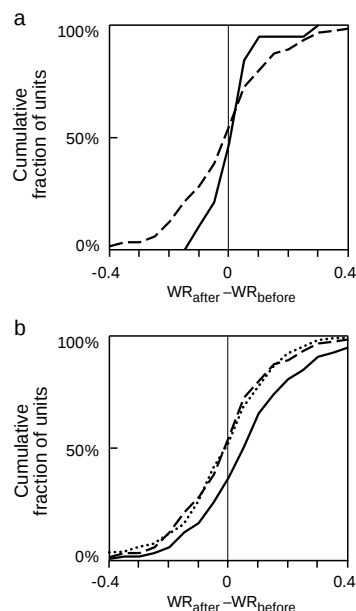


Figure 2 Statistical analysis of ACh-induced response modifications for all tested units. a, Cumulative distribution of response changes ($WR_{\text{after}} - WR_{\text{before}}$; see Methods for definition of weighted ratio (WR)) plotted for cells tested for temporal stability of TFTCs without ACh ($n = 19$, solid line) and with ACh ($n = 40$, dashed line). b, Cumulative distributions of changes between the second and fourth TFTCs. Changes in responses of the paired group ($n = 119$) to the paired (solid line) and the unpaired (dotted line) frequencies and in responses of the unpaired control group to all frequencies ($n = 40$, dashed line).

ness (87.5%). The overlap between the populations of cells displaying significant TFTC changes after pairing with and without ACh was small: only 3 out of the 22 modified single units showed effects in both conditions.

We tested whether ACh application is required during pairing in order to induce TFTC changes by repeating the pairing protocol without ACh. In most cases (17 out of 19), whisker stimulation without ACh induced no changes in the TFTC. Additional evidence supporting the permissive role of ACh in the induced plasticity was obtained from cells recorded simultaneously with a tungsten-in-glass electrode other than the combined electrode used for iontophoresis. Only 7% of the units (3 out of 45) recorded with tungsten-in-glass electrodes (that is, units that were probably beyond the maximal distance from the ejection site where ACh is still effective in modifying neuronal firing activity (S. H. *et al.*, unpublished data)) were modified after pairing when tested with ACh, even though they were activated by the stimulus. In contrast, 33% (39 out of 119) of the cells simultaneously recorded by the combined electrode were significantly modified (χ^2 test, $P < 0.002$), confirming that the modifications are observed preferentially within the region of ACh application.

Part of the response modifications described here may have resulted from the presence of ACh during the second and fourth TFTCs of the protocol and not from the pairing. To isolate the effect of the fixed-frequency pairing, the population of cells submitted to pairing was compared with a control population for which the four TFTCs were applied with no conditioning period in between (the time interval between the second and the third TFTCs was kept the same as for the original protocol). Responses were quantified by a weighted ratio between the response to stimulation at a given frequency and the averaged response to all other frequencies (see Methods). Figure 2 shows the cumulative distributions of changes

in weighted ratio observed in each condition. Consistent with other studies^{23,24}, and independently of the pairing, the response variability was larger (two-tailed F -test, $P < 0.001$) when tested with ACh (Fig. 2a, dashed line) than without ACh (Fig. 2a, solid line), whereas the mean was unchanged (two-tailed Mann-Whitney U -test, $P > 0.46$). However, the introduction of the fixed-frequency pairing induced an additional effect: the relative strength of the response to the paired frequency was significantly potentiated (Fig. 2b, compare solid and dashed curves, one-tailed Mann-Whitney U -test, $P < 0.001$), whereas the variability was unchanged (two-tailed F -test, $P > 0.37$). These potentiations were specific for the paired frequency, as the distribution of changes for unpaired frequencies (Fig. 2b, dotted line) was indistinguishable from the control distribution (Fig. 2b, dashed line, two-tailed Mann-Whitney U -test, $P > 0.1$), and significantly different from the distribution of changes at the paired frequency (Fig. 2b, solid line, two-tailed Mann-Whitney U -test, $P < 0.0001$).

Statistically, the entire population was potentiated by fixed-frequency pairings. However, when each unit was analysed separately, only a subpopulation exhibited significant modifications. We examined the dependency of these modifications on the frequency of the paired stimulus (Fig. 3). The significant potentiations observed during testing with ACh were maximal for the paired frequency (Fig. 3a–c) and differed from changes at other frequencies (Fig. 3d–f; analysis of variance (ANOVA), $F(1,43) = 5.45, 6.68$ and 6.43 for 5, 8 and 11 Hz, respectively, $P < 0.05$). On average, the TFTCs' reorganization after pairing was such that paired and unpaired frequencies showed, respectively, relative gains and losses in response (Fig. 3g, right box; ANOVA, $F(1,43) = 12.07$, $P < 0.005$). No significant reorganization of the TFTCs was observed when testing without ACh (Fig. 3g, left box; ANOVA, $F(1,43) = 0.45$, $P > 0.5$).

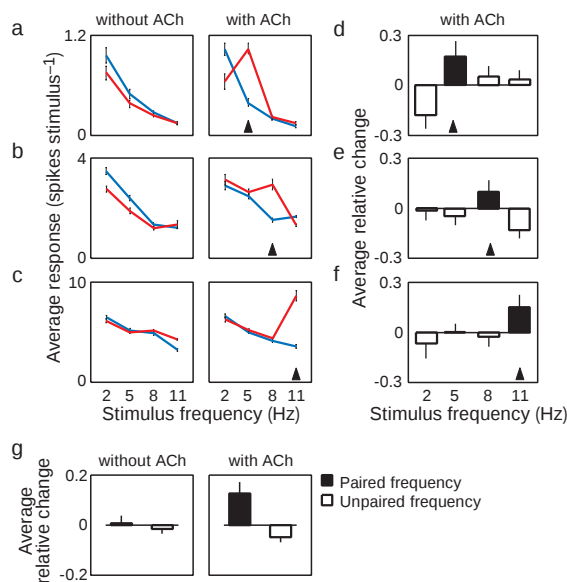


Figure 3 Reorganization of TFTCs expressed with ACh after pairing. **a–c**, Average responses ($\pm$ s.e.m.) to different stimulation frequencies. Three examples showing significant (Kolmogorov–Smirnov, $P < 0.0005$) response enhancements to the paired frequency (5, 8 or 11 Hz) with ACh (blue before, red after pairing, arrowheads indicate the paired frequency). TFTCs tested without ACh were unchanged ($P > 0.4$) (**a,b**) or showed smaller changes ($P < 0.002$) (**c**). **d–f**, For all units significantly potentiated at any frequency when tested with ACh, the response changes for each frequency were averaged across units submitted to pairings at 5 (**d**), 8 (**e**) or 11 (**f**) Hz. Response changes to the paired (black bars) and non-paired (white bars) frequencies differ within each group (ANOVA; $P < 0.05$). **g**, Changes in responses to paired and non-paired frequencies averaged across all potentiated cells tested with and without ACh.

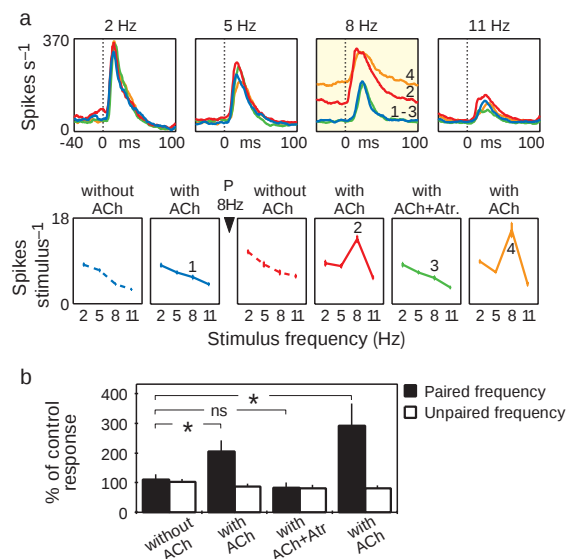


Figure 4 Atropine blocks the ACh-dependent expression of plasticity. **a**, Top row, peristimulus time histograms of responses before (blue), and after pairing tested with ACh (red), with ACh and atropine (green), and again with ACh (orange) are superimposed for each frequency. Bottom row, TFTCs (mean response $\pm$ s.e.m.) of the unit shown in the top row before (blue) and after (red, green, orange) pairing at 8 Hz (P). TFTCs obtained without ACh are shown as dashed lines. Numbers (1–4) indicate the response levels corresponding to the histograms numbered in the top row (8 Hz). **b**, Mean ($\pm$ s.e.m.) percentage of change in response to stimulation at paired (black bars) and unpaired (white bars) frequencies compared with control (in chronological order: tested without ACh, with ACh, with ACh and atropine, and again with ACh). Asterisk, two-tailed Student's t -test, $P < 0.05$, $n = 5$; ns, not significant.

The ACh-dependent expression of the enhancement in response to the paired frequency was blocked by the muscarinic antagonist, atropine. Figure 4 shows an example of a significant frequency-specific potentiation (Fig. 4a, 8 Hz, red line; Kolmogorov–Smirnov, $P < 10^{-7}$) that is absent when atropine and ACh are iontophoresed together during testing (Fig. 4a, green line; Kolmogorov–Smirnov, $P > 0.89$). Two minutes after the end of atropine application, a significantly enhanced response to the paired frequency was recovered with ACh (Fig. 4a, orange line; Kolmogorov–Smirnov, $P < 10^{-7}$). Overall, the muscarinic nature of the effect has been confirmed in all the cells showing an ACh-dependent expression of plasticity and tested with atropine (five out of five cells; Fig. 4b).

The temporal response properties of populations of auditory cortical cells can be modified after extensive periods of tone presentations at a given repetition rate paired with stimulation of the nucleus basalis¹³. We have shown that single units of the somatosensory barrel cortex can show a rapidly induced ACh-dependent plasticity of temporal response properties. Furthermore, we have shown that the expression of ACh-induced modifications is also regulated by increased cortical ACh. The altered responsiveness to a specific stimulus frequency, which was associated with increased ACh levels, was expressed only in the presence of ACh. The requisite for a similarity between the acquisition and the recall conditions is analogous to a “state-dependent learning”^{1,2}—a phenomenon in which newly acquired information may become available for retrieval only if the endogenous state of the brain and the sensory context present at the time of the original encoding episode are reinstated at the time of testing. In our anaesthetized animals, the increased cholinergic levels were induced by exogenous applications; however, in the awake animal, endogenous activation of the cholinergic system probably provides the required levels of cortical ACh for both memory formation^{5,8} and recall²⁵. □

Methods

Animal preparation and electrophysiology

Experiments were carried out on adult Wistar albino rats (300 ± 25 g) obtained from the Animal Breeding Unit of The Weizmann Institute of Science. Maintenance, manipulations and surgery were according to institutional animal welfare guidelines. Experimental procedures were similar to those used previously^{26,27}. Briefly, anaesthetized rats (urethane, 1.5 g kg⁻¹) were mounted in a modified stereotaxic device²⁸ which allows free access to the somatosensory cortex and to vibrissae. The right postero-medial barrel subfield was exposed, the dura removed and neural activity recorded with a multi-electrode array composed of two tungsten-in-glass electrodes and a combined electrode²⁹ composed of a tungsten core surrounded by six micropipettes. The pipettes were filled with acetylcholine chloride (1 M, pH 4.5), atropine sulphate (0.1 M, pH 4.5) and NaCl (3 M) for current balance. In most cases, the tungsten-in-glass and combined electrodes were lowered independently into different barrels. Data from units recorded by the combined electrodes ($n = 132$) and the tungsten-in-glass electrodes ($n = 52$) were analysed separately.

Vibrissae stimulation and protocol

Whiskers were stimulated by a linear electromagnetic vibrator (pulses of 10 ms, 5-ms rise time and 5-ms fall time, 160 μm at ~5 mm from the snout). Temporal frequency tuning curves (TFTCs) were obtained by deflecting the principal vibrissa at different frequencies in the following order: 2, 5, 8, 11 and in a few cases 14 Hz; 45 s interval; (14), 11, 8, 5, 2 Hz, with interblock intervals of 10 s. Stimuli were applied at each frequency in blocks of 12 consecutive trains of 4 s + 1 s intertrain interval each. Before pairing, the TFTC was determined first without and then during ACh iontophoresis. Pairing consisted of 24 trains of stimulation (each of 4 s + 1 s intertrain interval) of the vibrissa at one fixed temporal frequency (5, 8 or 11 Hz) accompanied with ACh iontophoresis (20–80 nA). After pairing, the TFTC was determined without ACh and once again with ACh. In some experiments ($n = 16$ cells), two additional TFTCs were determined, one during combined iontophoresis of ACh and atropine (60 nA) and another during ACh application alone. For 53 cells out of 119 only one pairing was applied. For the other recordings, the pairing was repeated several times at the same or different frequencies.

Data analysis

To keep the initial state comparable among cells, only the first paired frequency was considered for statistical tests. The effect was assessed systematically on the test period immediately after the last pairing at that frequency. The relative strength of the response to a given frequency was quantified by the weighted ratio (WR) = $(R_f - \text{AvgR}) / (R_f + \text{AvgR})$, where R_f is the response to stimulation at a given frequency (spike count over 60 ms from the stimulus onset) and AvgR is the averaged response to stimulation at all other frequencies. This ratio, which takes values from -1 to +1, was calculated for each of the 24

trains of stimuli, with R_f and AvgR values computed from corresponding trains across frequencies presented during the same TFTC. To assess the effect of conditioning, the 24 values obtained from TFTCs before and after pairing were statistically compared (two-tailed Kolmogorov–Smirnov, significance level $P < 0.01$). The comparison was performed independently for the TFTCs obtained without and with ACh, for each frequency. To assess the frequency specificity of the effect, cells were grouped as a function of the paired frequency (5, 8 or 11 Hz), and the differences in weighted ratios were averaged across cells (see Fig. 3d–f). This analysis was done on all cells showing a statistically significant change in weighted ratio values for any of the tested frequencies (paired and non-paired), thus avoiding any bias towards the paired frequency. The weighted values were statistically compared using multi-factor ANOVA with repeated measures.

Received 8 October; accepted 15 November 1999.

- Izquierdo, I. in *Neurobiology of Learning and Memory* (eds Lynch, G., McGaugh, J. L. & Weinberger, N. M.) 333–358 (Guilford, New York, 1984).
- Gordon, W. C. & Klein, R. L. in *Animal Learning and Cognition* (ed. Mackintosh, N. J.) 255–279 (Academic, San Diego, 1994).
- Mesulam, M. M., Mufson, E. J., Wainer, B. H. & Levey, A. I. Central cholinergic pathways in the rat: an overview based on an alternative nomenclature (Ch1–Ch6). *Neuroscience* **10**, 1185–1201 (1983).
- Richardson, R. T. & DeLong, M. R. A reappraisal of the functions of the nucleus basalis of Meynert. *Trends Neurosci.* **11**, 264–267 (1988).
- Singer, W. in *Brain Organization and Memory: Cells, Systems, and Circuits* (eds McGaugh, J. L., Weinberger, N. M. & Lynch, G.) 211–233 (Oxford Univ. Press, New York, 1990).
- Ahissar, E. & Ahissar, M. Plasticity in auditory cortical circuitry. *Curr. Opin. Neurobiol.* **4**, 580–587 (1994).
- Weinberger, N. M. Dynamic regulation of receptive fields and maps in the adult sensory cortex. *Annu. Rev. Neurosci.* **18**, 129–158 (1995).
- Dykes, R. Mechanisms controlling neuronal plasticity in somatosensory cortex. *Can. J. Physiol. Pharmacol.* **75**, 535–545 (1997).
- Edeline, J. M. Learning-induced physiological plasticity in the thalamo-cortical sensory systems: a critical evaluation of receptive field plasticity, map changes and their potential mechanisms. *Prog. Neurobiol.* **57**, 165–224 (1999).
- Metherate, R. & Weinberger, N. M. Acetylcholine produces stimulus-specific receptive field alterations in cat auditory cortex. *Brain Res.* **480**, 372–377 (1989).
- Edeline, J. -M., Hars, B., Maho, C. & Hennevin, E. Transient and prolonged facilitation of tone-evoked responses induced by basal forebrain stimulations in the rat auditory cortex. *Exp. Brain Res.* **97**, 373–386 (1994).
- Kilgard, M. P. & Merzenich, M. M. Cortical map reorganization enabled by nucleus basalis activity. *Science* **279**, 1714–1718 (1998).
- Kilgard, M. P. & Merzenich, M. M. Plasticity of temporal information processing in the primary auditory cortex. *Nature Neurosci.* **8**, 727–731 (1998).
- Rasmussen, D. D. & Dykes, R. W. Long-term enhancement of evoked potentials in cat somatosensory cortex produced by co-activation of the basal forebrain and cutaneous receptors. *Exp. Brain Res.* **70**, 276–286 (1988).
- Jacobs, S. E. & Juliano, S. L. The impact of basal forebrain lesions on the ability of rats to perform a sensory discrimination task involving barrel cortex. *J. Neurosci.* **15**, 1099–1109 (1995).
- Baskerville, K. A., Schweitzer, J. B. & Herron, P. Effects of cholinergic depletion on experience-dependent plasticity in the cortex of the rat. *Neurosci.* **80**, 1159–1169 (1997).
- Maalouf, M., Miasnikov, A. A. & Dykes, R. W. Blockade of cholinergic receptors in rat barrel cortex prevents long-term changes in the evoked potential during sensory preconditioning. *J. Neurophysiol.* **80**, 529–545 (1998).
- Sachdev, R. N., Lu, S. M., Wiley, R. G. & Ebner, F. F. Role of the basal forebrain cholinergic projection in somatosensory cortical plasticity. *J. Neurophysiol.* **79**, 3216–3228 (1998).
- Woolsey, T. A. & Van der Loos, H. The structural organization of layer IV in the somatosensory region (SI) of mouse cerebral cortex. The description of a cortical field composed of discrete cytoarchitectonic units. *Brain Res.* **17**, 205–242 (1970).
- Carvel, G. E. & Simons, D. J. Biometric analyses of vibrissal tactile discrimination in the rat. *J. Neurosci.* **10**, 2638–2648 (1990).
- Ahissar, E., Haidarliu, S. & Zacksenhouse, M. Decoding temporally encoded sensory input by cortical oscillations and thalamic phase comparators. *Proc. Natl Acad. Sci. USA* **94**, 11633–11638 (1997).
- Webster, H. H. et al. Long-term enhancement of evoked potentials in racoon somatosensory cortex following co-activation of the nucleus basalis of Meynert complex and cutaneous receptors. *Brain Res.* **545**, 292–296 (1991).
- Howard, M. A. & Simons, D. J. Physiologic effects of nucleus basalis magnocellularis stimulation on rat barrel cortex neurons. *Exp. Brain Res.* **102**, 21–33 (1994).
- Ahissar, E., Haidarliu, S. & Shulz, D. E. Possible involvement of neuromodulatory systems in cortical Hebbian-like plasticity. *J. Physiol. (Paris)* **90**, 353–360 (1996).
- Delacour, J., Houcine, O. & Costa, J. C. Evidence for a cholinergic mechanism of “learned” changes in the responses of barrel field neurons of the awake and unanesthetized rat. *Neuroscience* **34**, 1–8 (1990).
- Shulz, D. E., Cohen, S., Haidarliu, S. & Ahissar, E. Differential effects of acetylcholine on neuronal activity and interactions in the auditory cortex of the guinea-pig. *Eur. J. Neurosci.* **9**, 396–409 (1997).
- Haidarliu, S. & Ahissar, E. Spatial organization of facial vibrissae and cortical barrels in the guinea pig and golden hamster. *J. Comp. Neurol.* **385**, 515–527 (1997).
- Haidarliu, S. An anatomically adapted, injury-free headholder for guinea pigs. *Physiol. Behav.* **60**, 111–114 (1996).
- Haidarliu, S., Shulz, D. E. & Ahissar, E. A multi-electrode array for combined microiontophoresis and multiple single-unit recordings. *J. Neurosci. Meth.* **56**, 125–131 (1995).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank Y. Dudai, Y. Frégnac, H. Markram, P. Salin and M. Segal for helpful comments, and A. Bagady for help with statistical analysis. V.E. and D.E.S. were supported by the

French Embassy in Israel during their visits to the laboratory of E.A. where the experiments were done. This work was supported by PICS CNRS, Ministère des Affaires Étrangères Français, AFIRST, HFSP, US-Israel Binational Science Foundation, Israel and the MINERVA Foundation, Germany.

Correspondence and requests for materials should be addressed to D.E.S. (e-mail: shulz@iaf.cnrs-gif.fr).

Modulation of A-type potassium channels by a family of calcium sensors

W. Frank An^{*}, Mark R. Bowlby[†], Maria Betty[†], Jie Cao^{*}, Huai-Ping Ling[†], Grace Mendoza[†], Joseph W. Hinson[†], Karen I. Mattsson[†], Brian W. Strassle[†], James S. Trimmer[‡] & Kenneth J. Rhodes[†]

^{*}Millennium Pharmaceuticals, Cambridge, Massachusetts 02139, USA

[†]Neuroscience Division, Wyeth-Ayerst Research, Princeton, New Jersey 08543, USA

[‡]Department of Biochemistry and Cell Biology, SUNY, Stony Brook, New York 11794, USA

In the brain and heart, rapidly inactivating (A-type) voltage-gated potassium (Kv) currents operate at subthreshold membrane potentials to control the excitability of neurons and cardiac myocytes^{1,2}. Although pore-forming α -subunits of the Kv4, or *Shal*-related, channel family form A-type currents in heterologous cells³, these differ significantly from native A-type currents. Here we describe three Kv channel-interacting proteins (KChIPs) that bind to the cytoplasmic amino termini of Kv4 α -subunits. We find that expression of KChIP and Kv4 together reconstitutes several features of native A-type currents by modulating the density, inactivation kinetics and rate of recovery from inactivation of Kv4 channels in heterologous cells. All three KChIPs co-localize and co-immunoprecipitate with brain Kv4 α -subunits, and are thus integral components of native Kv4 channel complexes. The

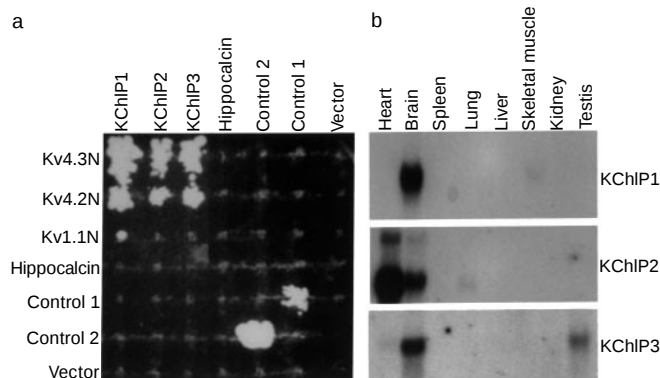


Figure 1 YTH interaction and tissue expression of KChIPs 1–3. **a**, Growth on the seventh day on a Trp-Leu-His synthetic dropout plate (with 10 mM 3-AT) demonstrating the specific interaction of KChIPs 1–3 with Kv4 α -subunits. KChIP1, KChIP2, KChIP3 and hippocalcin are in the 'fish' configuration (columns); N-terminal domains of Kv4.3, Kv4.2, Kv1.1 and hippocalcin are in the 'bait' configuration (rows). Control 1 and control 2 are two interacting fish/bait pairs, respectively, unrelated to K⁺ channels or the KChIPs. **b**, Northern blots (Clontech) showing the expression of KChIP mRNAs across rat (KChIP1 and KChIP2) and mouse (KChIP3) tissues.

KChIPs have four EF-hand-like domains and bind calcium ions. As the activity and density of neuronal A-type currents tightly control responses to excitatory synaptic inputs, these KChIPs may regulate A-type currents, and hence neuronal excitability, in response to changes in intracellular calcium.

A-type Kv currents formed by Kv4-family α -subunits control excitatory responses in neuronal cell bodies and dendrites^{1,3–7} and contribute to repolarization following action potentials in cardiac myocytes². Here we used the yeast two-hybrid (YTH) system⁸ to identify proteins that modulate Kv4 channels. We constructed a YTH bait corresponding to the intracellular amino terminus (amino acids 1–180) of the rat Kv4.3 subunit and screened an oligo dT-primed library of rat midbrain complementary DNA to identify proteins that interacted with it. Many proteins that strongly interacted with the Kv4.3 N-terminal bait also interacted with the N-terminal 180 amino acids of Kv4.2, but not with Kv1.1 or other, unrelated baits (Fig. 1a). Among the Kv4-specific interactors were two new members of a previously described gene family (see below), here termed KChIP1 and KChIP2. Library screening and database mining identified mouse and human orthologues of these genes, as well as expressed sequence tags (ESTs) encoding a previously identified member of this family (KChIP3). Northern blot analysis of rat (KChIP1 and KChIP2) or mouse (KChIP3) tissues revealed that KChIP1 is predominantly expressed in brain, KChIP2 is expressed in heart, brain and lung, and KChIP3 is highly expressed in brain with lower expression in testes (Fig. 1b).

The KChIP1, 2 and 3 cDNAs encode 216-, 252- and 256-amino-acid polypeptides, respectively, which have distinct N termini but share ~70% amino-acid identity throughout a carboxy-terminal 185-amino-acid 'core' domain containing four EF-hand-like motifs (Fig. 2). Although these KChIPs have around 40% amino-acid similarity to neuronal calcium sensor-1 (NCS-1) and are members of the recoverin/NCS subfamily of calcium-binding proteins⁹ (Fig. 2), other members of this subfamily (such as hippocalcin) did not

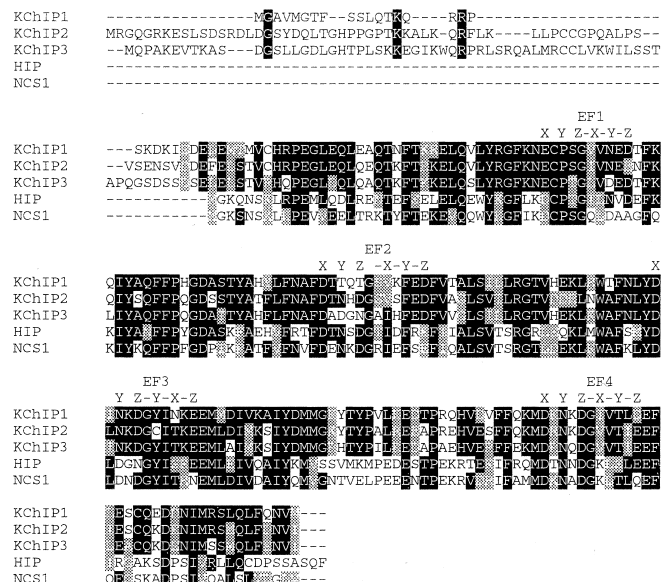


Figure 2 Sequence alignment of human KChIPs with members of the recoverin family of Ca²⁺-sensing proteins. The alignment was performed using CLUSTALW²¹. Residues identical to the consensus are shaded black and conservative substitutions are shaded grey. X marks position 1 of the 12-amino-acid consensus EF-hand motif, as defined in PROSITE²². X, Y, Z and -X, -Y, -Z denote EF-hand Ca²⁺-binding residues. Like all members of the recoverin family, the KChIP EF1 diverges significantly from the EF-hand consensus and contains a Cys-Pro motif. HIP, human hippocalcin; NCS1, rat neuronal calcium sensor-1.

interact with Kv4 channels in the YTH assay (Fig. 1a). These KChIPs also share sequence similarity with frequenin, a *Drosophila* calcium-binding protein that modulates ion channels in the plasma membrane¹⁰. Moreover, KChIP3 is identical to calsenilin¹¹, and 99% identical at the nucleotide level to the Ca²⁺-regulated transcriptional modulator DREAM¹², indicating that these KChIPs may have activity beyond modulation of Kv4 channels.

To confirm that the KChIP polypeptides associate with expressed, full-length Kv4 α -subunits, the KChIP1 cDNA was transiently expressed alone or with Kv4.2 in COS-1 and Chinese hamster ovary (CHO) cells (Fig. 3). We monitored expression of KChIP1 and Kv4.2 by immunofluorescent staining using mouse monoclonal or affinity-purified rabbit polyclonal antibodies raised against recombinant KChIP and Kv4 polypeptides. When expressed alone, KChIP1 is diffusely distributed throughout the cytoplasm of COS-1 cells (Fig. 3a). In contrast, Kv4.2 is concentrated within the perinuclear endoplasmic reticulum and Golgi compartments, with some immunoreactivity apparent at the outer margins of the cell (Fig. 3b). When KChIP1 is expressed with Kv4.2, the characteristic diffuse KChIP1 distribution changes markedly, such that KChIP1 and Kv4.2 immunofluorescence completely overlap (Fig. 3c). This redistribution of KChIP1 immunofluorescence did not occur when KChIP1 was expressed with the Kv1.4 α -subunit (data not shown), indicating that altered KChIP1 localization is not a consequence of overexpression and that KChIP1 associates preferentially with Kv4-family α -subunits. Identical results were obtained when KChIP2 or KChIP3 were expressed with Kv4.2 or Kv4.3, indicating that all three KChIPs co-localize with Kv4 α -subunits in transiently transfected COS-1 cells. Reciprocal immunoprecipitation analyses using KChIP and channel-specific antibodies revealed that the KChIPs associate tightly with Kv4 α -subunits in co-transfected cells. For example, anti-Kv4.2 antibodies co-immunoprecipitate KChIP1, and anti-KChIP1 antibodies co-immunoprecipitate Kv4.2, from co-transfected cell lysates (Fig. 4a).

Next we examined the localization and association of these KChIPs and Kv4 subunits in adult rat brain. Immunohistochemical staining of rat brain sections revealed that KChIP1 and KChIP2

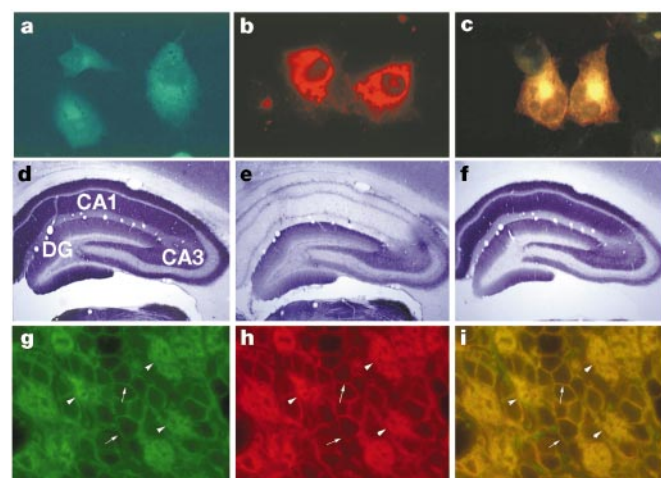


Figure 3 Immunolocalization of Kv4 and KChIP1 in COS-1 cells and rat brain. **a–c**, COS-1 cells were transiently transfected with KChIP1 cDNA alone (**a**), Kv4.2 alone (**b**) or KChIP1 and Kv4.2 (**c**). In co-transfected cells, KChIP1 co-localizes with Kv4.2. **d–f**, Sections of rat hippocampus were stained¹⁹ to visualize Kv4.2 (**d**), Kv4.3 (**e**) or KChIP2 (**f**). In the dendritic fields of the dentate gyrus (DG) and CA3 subfield, Kv4.2 and Kv4.3 co-localize with KChIP2. **g–i**, Cryosections of adult rat cerebellum were double-labelled¹⁹ for KChIP1 (**g**) and Kv4.2 (**h**). As shown in the merged image (**i**), Kv4.2 and KChIP1 co-localize in synaptic glomeruli (arrowheads) and in the somatodendritic membranes of granule cells (arrows).

co-localize with Kv4.2 and Kv4.3 in a region- and cell-type-specific manner (Fig. 3). For example, KChIP2 co-localizes with Kv4.2 in the dendrites of granule cells in the dentate gyrus (Fig. 3d–f), in the apical and basal dendrites of hippocampal and neocortical pyramidal cells, and in several subcortical structures including the striatum and thalamus. KChIP1 co-localizes with Kv4.3 in hippocampal interneurons, cerebellar granule cells and synaptic glomeruli within the cerebellar granule cell layer (Fig. 3g–i). Co-immunoprecipitation analyses confirmed that the KChIPs are tightly associated

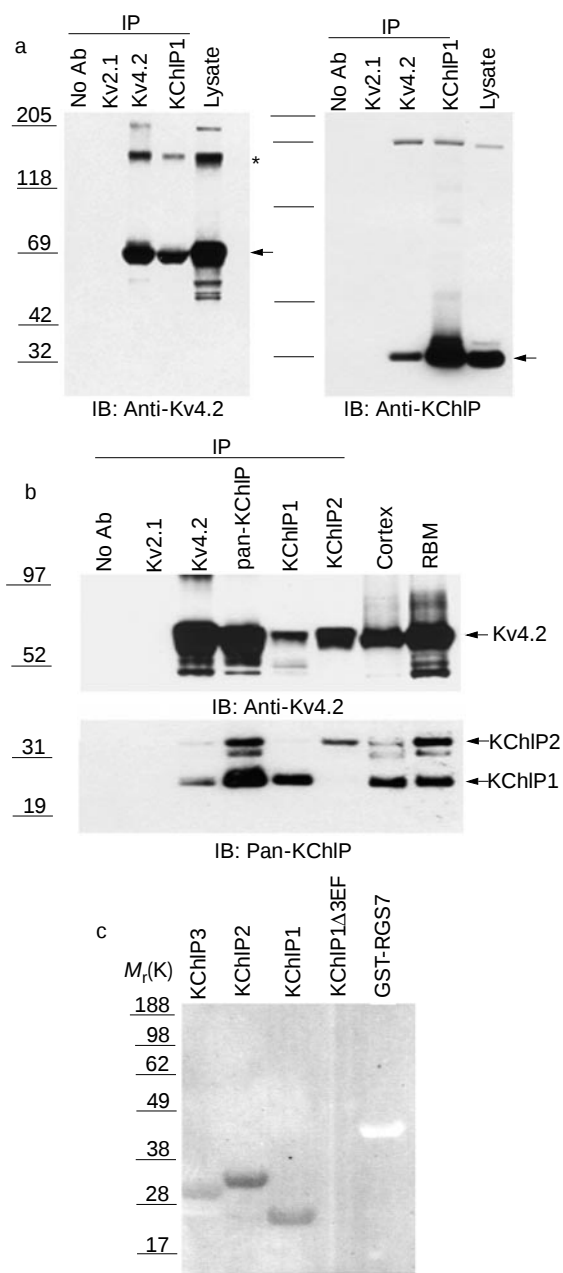


Figure 4 KChIPs co-associate with Kv4 α -subunits. **a**, Lysates of Kv4.2/KChIP1 co-transfected CHO cells were immunoprecipitated using Kv4.2 and KChIP-specific antibodies. The immunoprecipitation products were separated by 4–12% and 10% SDS-PAGE, (left and right panels, respectively) and the resulting immunoblots were probed using Kv4.2- and KChIP1-specific antibodies. Asterisk, aggregated Kv4.2. **b**, Rat neocortex membranes were immunoprecipitated as described. The resulting immunoblots were probed with anti-Kv4.2 or 'pan-KChIP' antibodies. **c**, Recombinant KChIPs were separated by 4–20% SDS-PAGE, transferred to nitrocellulose and probed with [⁴⁵Ca²⁺]¹⁴.

with Kv4 α -subunits in brain potassium channel complexes. As shown in Fig. 4b, anti-KChIP antibodies efficiently co-immunoprecipitate Kv4.2 from rat neocortical membranes, and anti-Kv4.2 antibodies efficiently co-immunoprecipitate KChIP1 and 2. The KChIP polypeptides were not co-immunoprecipitated with Kv2.1, indicating that these KChIPs may preferentially associate with Kv4 α -subunits in rat brain membranes. Taken together, our immunohistochemical and immunoprecipitation results confirm that the KChIPs are integral components of native Kv4 channel complexes.

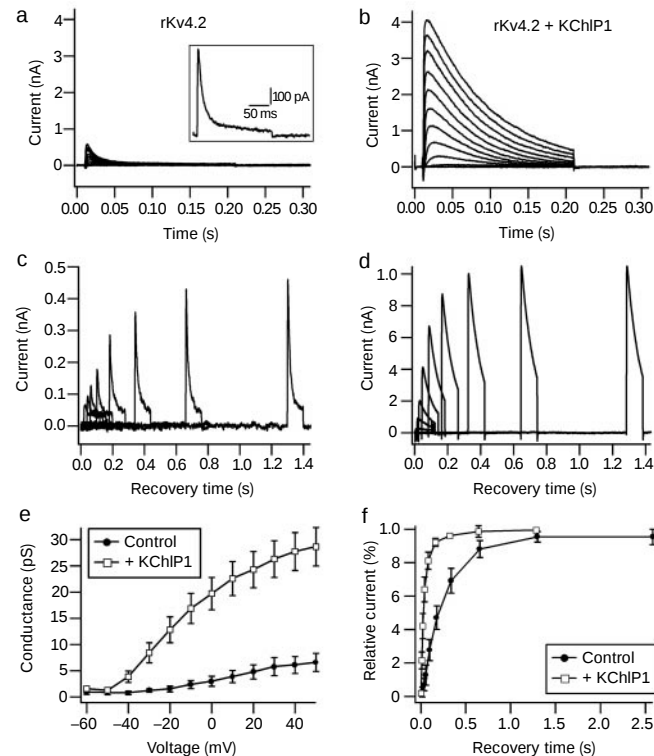


Figure 5 KChIP1 alters the density and kinetics of Kv4.2 currents in CHO cells. **a, b**, Current–voltage traces from CHO cells transiently transfected with Kv4.2 (**a**) or Kv4.2 and KChIP1 (**b**). Inset, single current trace at 0 mV shown on expanded axes. **c, d**, Recovery from inactivation using a two-pulse protocol (first pulse not shown, see Methods) for Kv4.2 alone (**c**) or Kv4.2 and KChIP1 (**d**). **e**, Peak current conductance at all voltages from cells transfected with Kv4.2 alone ($n = 11$) or co-transfected with KChIP1 ($n = 9$). **f**, Summary of the percentage of peak current recovery from inactivation for Kv4.2 ($n = 8$) and Kv4.2/KChIP1 ($n = 5$) co-transfected cells.

As the KChIPs localize and associate with native Kv4 α -subunits, we next determined whether they modulate the electrophysiological properties of expressed Kv4 channels. Transient transfection of the rat Kv4.2 α -subunit in CHO cells yielded typical A-type potassium currents^{3,13}. Expression of Kv4.2 together with KChIP1, 2 or 3 revealed several effects of KChIP co-expression on Kv4 currents (Fig. 5 and Table 1). First, the density of Kv4.2 currents increased about 12-fold (Kv4.2 alone, 25.5 ± 3.2 pA pF⁻¹; Kv4.2 + KChIP1, 306.9 ± 57.9 pA pF⁻¹), indicating that KChIP1 may promote and/or stabilize expression of Kv4.2 at the cell surface. Second, the mid-point of voltage activation (activation $V_{1/2}$) of Kv4.2 currents shifted to more hyperpolarized potentials (activation $V_{1/2}$ for Kv4.2 alone, 28.8 ± 7.0 mV; Kv4.2 + KChIP1, -12.1 ± 1.4 mV). Furthermore, the kinetics of Kv4.2 inactivation slowed considerably (inactivation time constant of Kv4.2 alone, 28.2 ± 2.6 ms; Kv4.2 + KChIP1, 104.1 ± 10.4 ms), whereas KChIP/Kv4.2 channels recovered from inactivation much more rapidly (KChIP1/Kv4.2 recovery $\tau = 53.6 \pm 7.6$ ms) versus channels produced by Kv4.2 expression alone (recovery $\tau = 272.2 \pm 26.1$ ms).

To investigate the role of the KChIPs' unique N termini in modulating Kv4 channels, we truncated KChIP1 and KChIP2 to the highly conserved C-terminal 185-amino-acid core sequence. The effects of KChIP1 Δ N2-31 or KChIP2 Δ N2-67 on Kv4.2 current density and kinetics were very similar to those produced by the full-length KChIPs (Table 1), indicating that the N termini do not contribute significantly to modulation of Kv4 channels. To determine whether the modulatory effects of these KChIPs are specific for the Kv4 channels among the broader Kv channel family, KChIP1 was expressed along with Kv1.4 or Kv2.1 in *Xenopus* oocytes. KChIP1 did not modulate Kv1.4 or Kv2.1 currents (Table 2), indicating that these KChIPs preferentially modulate Kv4 channels. To verify that

Table 2 Functional effects of KChIPs on other Kv channels

Current parameter	Oocytes		Oocytes	
	hKv1.4	hKv1.4 + KChIP1	hKv2.1	hKv2.1 + KChIP1
Peak current (μ A per cell at 50 mV)	8.3 $\pm$ 2.0	6.5 $\pm$ 0.64	3.7 $\pm$ 0.48	2.9 $\pm$ 0.37
Inactivation time constant (ms at 50 mV)	53.2 $\pm$ 2.8	58.2 $\pm$ 6.6	1.9 s $\pm$ 0.079	1.7 s $\pm$ 0.078
Recovery from inactivation time constant (s, at -80 mV)	1.9	1.6	7.6	7.7
Activation $V_{1/2}$ (mV)	-21.0	-20.9	12.0	12.4
Steady-state inactivation $V_{1/2}$ (mV)	-48.1	-47.5	-25.3	-23.9
Number of cells (n)	4	4	5	7

Table 1 Functional effects of KChIPs on Kv4 channels

Current parameter	CHO						Oocytes	
	rKv4.2 + vector	rKv4.2 + KChIP1	rKv4.2 + KChIP1 Δ N2-31	rKv4.2 + KChIP1 EF-mutant	rKv4.2 + KChIP2	rKv4.2 + KChIP2 Δ N2-67	hKv4.3 + KChIP3	hKv4.3 + KChIP1
Peak current (nA per cell at 50 mV)	0.60 $\pm$ 0.096	4.5* $\pm$ 0.55	6.0* $\pm$ 1.1	0.49 $\pm$ 0.22	3.3* $\pm$ 0.45	5.8* $\pm$ 1.1	3.5* $\pm$ 0.99	7.7 μ A $\pm$ 2.6
Peak current density (pA pF ⁻¹ at 50 mV)	22.5 $\pm$ 3.2	306.9* $\pm$ 57.9	407.2* $\pm$ 104.8	16.8 $\pm$ 4.3	196.6* $\pm$ 26.6	202.6* $\pm$ 27.5	161.7* $\pm$ 21.8	—
Inactivation time constant (ms at 50 mV)	28.2 $\pm$ 2.6	104.1* $\pm$ 10.4	129.2* $\pm$ 14.2	19.0* $\pm$ 2.7	95.1* $\pm$ 8.3	109.5* $\pm$ 9.6	67.2* $\pm$ 14.1	56.3 $\pm$ 6.6
Recover from inactivation (τ in ms at -80 mV)	272.2 $\pm$ 26.1	53.6* $\pm$ 7.6	98.1* $\pm$ 8.3	—	49.5* $\pm$ 9.0	36.1* $\pm$ 4.4	126.1* $\pm$ 16.2	327.0 $\pm$ 21.6
Activation $V_{1/2}$ (mV)	28.8 $\pm$ 7.0	-12.1* $\pm$ 1.4	-6.6* $\pm$ 1.9	14.6 $\pm$ 5.4	-1.2* $\pm$ 1.3	0.11* $\pm$ 1.8	-9.8* $\pm$ 1.7	-21.0 $\pm$ 3.1
Number of cells (n)	37	27	9	7	19	12	7	5

* Significantly different from control ($P < 0.05$).

the KChIPs' effects on Kv4 currents are independent of the expression system, the kinetic analyses described above were repeated after expressing Kv4.3 and KChIP in *Xenopus* oocytes. The effects of KChIP1 on Kv4.3 in the oocyte system were nearly identical to its effects on Kv4.2 in CHO cells (Table 1).

All three KChIPs, but not an unrelated glutathione S-transferase (GST)-fusion protein, bound $^{45}\text{Ca}^{2+}$ in a filter overlay assay¹⁴ (Fig. 4c) and displayed calcium-dependent mobility shifts on SDS-PAGE (data not shown), confirming that KChIP1, 2 and 3 are calcium-binding proteins. To investigate the functional significance of this calcium binding, we mutated the highly conserved Asp and Gly residues at positions 1 and 6, respectively, to Ala within EF-hands 2, 3 and 4 of KChIP1 (ref. 15). This KChIP1 triple EF-hand mutant does not bind $^{45}\text{Ca}^{2+}$ in the filter overlay assay (Fig. 4c), but does co-localize and is efficiently co-immunoprecipitated with Kv4 α -subunits upon co-transfection in COS-1 cells (data not shown). However, this mutant does not modulate Kv4.2 currents (Table 1), indicating that the binding interaction between KChIP1 and Kv4.2 is independent of calcium or tolerates EF-hand mutations, whereas modulation of Kv4.2 kinetics is dependent on calcium or is highly sensitive to point mutations within the EF-hand domains.

In summary, the effects of KChIP1, 2 and 3 on Kv4 channels indicate that, in native cells, these proteins may augment A-type currents by a variety of mechanisms: increased surface channel density, shifting the activation $V_{1/2}$ to more negative potentials, slower inactivation and faster recovery from inactivation. These effects of the KChIPs are very similar to those produced by unidentified factors encoded in low molecular weight (LMW) brain mRNA^{16,17}, although the KChIPs slowed inactivation, whereas factors in LMW RNA make Kv4 inactivation faster¹⁶. Although there may be additional modulatory components of Kv4 channel complexes, our data indicate that these KChIPs are integral regulatory components of Kv4 channels and that expression of these KChIPs with Kv4 α -subunits yields A-type currents with properties that more closely resemble those recorded from mammalian central neurons^{1,4,7}. The association of these KChIPs with Kv4 α -subunits in brain, coupled with their ability to bind calcium, indicates that the KChIPs may dynamically regulate dendritic excitability, coupling A-current density to activity-dependant fluctuations in intracellular calcium. □

Methods

Yeast two-hybrid screen

The YTH cDNA library was constructed from poly(A)⁺ RNA extracted from rat midbrain and ligated into the pACT2 vector (Stratagene). About 6 million yeast transformants were screened on Trp-Leu-His dropout plates in the presence of 10 mM 3-amino triazole (3-AT). His⁺ clones were assayed for β -galactosidase activity and His⁺/LacZ⁺ clones were rescued and sequenced. Mating assays were done essentially as described¹⁸. Growth was followed for up to seven days in the presence of various concentrations of 3-AT.

Immunoblotting and immunoprecipitation

Immunoblotting and immunoprecipitation analyses were performed as described¹⁹. The anti-KChIP and anti-Kv4 antibodies were determined to be specific by demonstrating positive immunofluorescence in COS-1 cells transfected with the cognate but not other cDNAs and immunoreactivity on immunoblots of lysates prepared from COS-1 cells transiently transfected with the cognate cDNA but not other cDNAs, and by confirming that the pattern of immunostaining in rat brain tissue sections was consistent with the pattern of mRNA expression revealed by *in situ* hybridization histochemistry (not shown). A detailed characterization of these antibodies will be published elsewhere.

Immunohistochemistry

Immunoperoxidase and immunofluorescent labelling of paraformaldehyde fixed rat brain sections was performed as described¹⁹. Double-label immunofluorescence staining of rat cerebellum was done on 1- μm -thick cryosections.

Electrophysiology

Channel and KChIP cDNAs in the pRBG4 vector for expression in mammalian cells, and the SP64 or pBluescript.KSM vector for translation and expression in *Xenopus laevis* oocytes, were expressed transiently using cationic lipids for CHO transfections and microinjection of *in vitro* transcribed mRNA for oocytes²⁰. Equal amounts of cDNA or

mRNA were used for KChIP/Kv4 co-expression analyses. Ionic currents were recorded using the whole-cell mode of the patch-clamp technique in CHO cells and with two-electrode voltage clamp in oocytes. Current-voltage curves were evoked by depolarizing cells from a holding potential of -80 mV to test potentials from -60 to 50 mV, and leak subtracted using a *p/5* protocol. Recovery from inactivation was measured by driving the channels into an inactivated state with a pulse to 50 mV, then hyperpolarizing the cells and applying a second pulse to 50 mV at various intervals (holding potential is -80 mV between pulses). Data was collected using an EPC9 amplifier and Pulse software (HEKA Elektronik) and analysed using Pulse, PulseFit and SAS software. Inactivation and recovery from inactivation were fit to a single exponential equation and the activation $V_{1/2}$ was calculated by fitting the conductance-voltage relationship for each cell to a Boltzmann function. We used the following solutions for recordings (in mM): CHO bath: Hanks Balanced Salt Solution (GibcoBRL), 10 dextrose, 10 Hepes, pH 7.4; pipette: 140 KCl, 2 MgCl₂, 10 Hepes, 10 EGTA, pH 7.4; oocyte-bath: 96 NaCl, 2 KCl, 1 CaCl₂, 1 MgCl₂, 10 Hepes, 50 $\mu\text{g ml}^{-1}$ Gentamycin, pH 7.6; electrode: 3M KCl. Constructs are always compared with vector controls and/or KChIP1 parallel co-transfections; however, for display purposes data for vector and KChIP1 controls were averaged across several different experiments. Statistical comparisons were made using the Z test with the Bonferroni correction for multiple comparisons.

Received 25 August; accepted 24 November 1999.

- Hoffman, D. A., Magee, J. C., Colbert, C. M. & Johnston, D. K⁺ channel regulation of signal propagation in dendrites of hippocampal pyramidal neurons. *Nature* **387**, 869–875 (1997).
- Dixon, J. E. et al. Role of the Kv4.3 K⁺ channel in ventricular muscle. A molecular correlate for the transient outward current. *Circ. Res.* **79**, 659–668 (1996).
- Seródió, P., Kentros, C. & Rudy, B. Identification of molecular components of A-type channels activating at subthreshold potentials. *J. Neurophys.* **72**, 1516–1529 (1994).
- Hoffman, D. A., & Johnston, D. Downregulation of transient K⁺ channels in dendrites of hippocampal CA1 pyramidal neurons by activation of PKA and PKC. *J. Neurosci.* **18**, 3521–3528 (1998).
- Sheng, M., Tsaur, M. L., Jan, Y. N. & Jan, L. Y. Subcellular segregation of two A-type K⁺ channel proteins in rat central neurons. *Neuron* **9**, 271–284 (1992).
- Seródió, P., Vega-Saenz de Miera, E. & Rudy, B. Cloning of a novel component of A-type K⁺ channels operating at subthreshold potentials with unique expression in heart and brain. *J. Neurophys.* **75**, 2174–2179 (1996).
- Song, W. J. et al. Somatodendritic depolarization-activated potassium currents in rat neostriatal cholinergic interneurons are predominantly of the A type and attributable to coexpression of Kv4.2 and Kv4.1 subunits. *J. Neurosci.* **18**, 3124–3137 (1998).
- Fields, S. & Song, O. A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245–246 (1989).
- Nef, P. in *Guidebook to the Calcium-Binding Proteins* (ed. Celio, M. R.) 94–97 (Oxford Univ. Press, New York, 1996).
- Pongs, O. et al. Frequentin—a novel calcium-binding protein that modulates synaptic efficacy in the *Drosophila* nervous system. *Neuron* **11**, 15–28 (1993).
- Buxbaum, J. D. et al. Calsenilin: a calcium-binding protein that interacts with the presenilins and regulates the levels of a presenilin fragment. *Nature Med.* **4**, 1177–1181 (1998).
- Carrion, A. M., Link, W. A., Ledo, F., Mellstrom, B. & Naranjo, J. R. DREAM is a Ca²⁺-regulated transcriptional repressor. *Nature* **398**, 80–84 (1999).
- Pak, M. D. et al. mShal, a subfamily of A-type K⁺ channel cloned from mammalian brain. *Proc. Natl Acad. Sci. USA* **88**, 4386–4390 (1991).
- Kobayashi, M., Takamatsu, K., Saitoh, S. & Noguchi, T. Myristoylation of hippocalcin is linked to its calcium-dependent membrane association properties. *J. Biol. Chem.* **268**, 18898–18904 (1993).
- Linse, S. & Forsen, S. in *Advances in Second Messenger and Phosphoprotein Research* (ed. Means, S.) 89–150 (Raven, New York, 1995).
- Chabala, L. D., Bakry, N. & Covarrubias, M. Low molecular weight poly(A)⁺ mRNA species encode factors that modulate gating of a non-Shaker A-type K⁺ channel. *J. Gen. Physiol.* **102**, 713–728 (1993).
- Rudy, B., Hoyer, J. H., Lester, H. A. & Davidson, N. At least two mRNA species contribute to the properties of rat brain A-type potassium channels expressed in *Xenopus* oocytes. *Neuron* **1**, 649–658 (1988).
- Finley, R. L. & Brent, R. Interaction mating reveals binary and ternary connections between *Drosophila* cell cycle regulators. *Proc. Natl Acad. Sci. USA* **91**, 12980–12984 (1994).
- Rhodes, K. J. et al. Association and colocalization of the KvB1 and KvB2 β -subunits with Kv1 α -subunits in mammalian brain K⁺ channel complexes. *J. Neurosci.* **17**, 8246–8258 (1997).
- Bowlby, M. R. & Levitan, I. B. Block of cloned voltage-gated potassium channels by the second messenger diacylglycerol independent of protein kinase C. *J. Neurophys.* **73**, 2221–2229 (1995).
- Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).
- Bairoch, A. & Cox, J. A. EF-hand motifs in inositol phospholipid-specific phospholipase C. *FEBS Lett.* **269**, 454–456 (1990).

Acknowledgements

We thank P. Chanda and W. Edris for generation and purification of the recombinant KChIP proteins; J. Wardwell-Swanson and S. Nawoschik for generating mammalian and oocyte expression constructs; L. Buchwalder for preparation of the anti-KChIP and anti-Kv4 mouse monoclonal antibodies; J. Tang, K. Maden and S. Dembski for help with the yeast two-hybrid screen; K. Young, Q. Wang, Y. M. Xie, T. Novak, C. Gimeno and P. Errada for technical advice; and J. Moyer, J. Barrett, M. Cockett, P. McGonigle, J. Lillie, R. Breitbart, K. Willis and P. DiStefano for support and encouragement.

Correspondence and requests for materials should be addressed to K.J.R. (e-mail: rhodesk2@war.wyeth.com). The KChIP cDNAs have been assigned the following GenBank accession numbers: KChIP1, AF199597; KChIP2, AF199598; KChIP3, AF199599.

Slowed recovery of rod photoresponse in mice lacking the GTPase accelerating protein RGS9-1

Ching-Kang Chen*, Marie E. Burns†, Wei He‡, Theodore G. Wensel‡, Denis A. Baylor† & Melvin I. Simon*§

* Division of Biology, 147-75, California Institute of Technology, Pasadena, California 91125, USA

† Department of Neurobiology, Stanford University Medical Center, Stanford, California 94305, USA

‡ Verna and Marrs McLean Department of Biochemistry, Baylor College of Medicine, Houston, Texas 77030, USA

Timely deactivation of the α -subunit of the rod G-protein transducin (G α t) is essential for the temporal resolution of rod vision¹. Regulators of G-protein signalling (RGS) proteins accelerate hydrolysis of GTP by the α -subunits of heterotrimeric G proteins²⁻⁴ *in vitro*. Several retinal RGS proteins can act *in vitro*

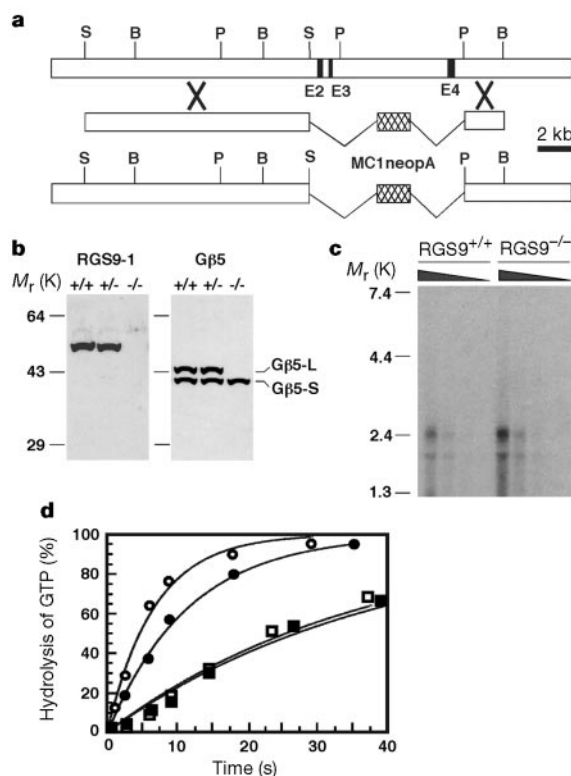


Figure 1 Inactivation of the RGS9 gene results in the loss of RGS9-1 and Gβ5-L in the retina and a reduced rate of hydrolysis of GTP by transducin. **a**, Inactivation of the RGS9 gene. Top, Partial genomic structure of the RGS9 gene; middle, targeting construct; bottom, altered RGS9 locus. The targeting construct contained 9 kb (long arm) of the 3' downstream and 1.3 kb (short arm) of the 5' upstream homologous sequence. The MC1neopA cassette (hatched box) replaced a 5.5-kb fragment that contained three exons corresponding to amino-acid residues 19–105. Restriction sites: B, *Bam*HI; P, *Pst*I, S, *Spe*I. **b**, Immunoblot analysis of RGS9-1 and Gβ5 protein levels in retinal homogenates derived from mice at 1 month old. The levels of RGS9-1 and Gβ5 were similar in RGS9^{+/+} and RGS9^{-/-} retinas. **c**, Northern blot analysis of Gβ5-L mRNA. From left to right, 4, 2, 1 and 0.5 μ g of total retinal RNA derived from mice at 1 month of age were loaded per lane. The retina-specific exon of the Gβ5 gene¹¹ was used as a probe. **d**, Single-turnover GTP hydrolysis by Gαt in ROS from RGS9^{+/+} (circles) and RGS9^{-/-} (squares) mice with (open) or without (filled) exogenous PDE-γ (1.33 μ M). Rate constants were determined by fitting the results with a single exponential function and are given in the text.

as GTPase accelerating proteins (GAP) for Gαt⁵⁻⁸. Recent reconstitution experiments indicate that one of these, RGS9-1, may account for much of the Gαt GAP activity in rod outer segments (ROS)^{8,9}. Here we report that ROS membranes from mice lacking RGS9-1 hydrolyse GTP more slowly than ROS membranes from control mice. The Gβ5-L protein that forms a complex with RGS9-1 (ref. 10) was absent from RGS9^{-/-} retinas, although Gβ5-L messenger RNA was still present. The flash responses of RGS9^{-/-} rods rose normally, but recovered much more slowly than normal. We conclude that RGS9-1, probably in a complex with Gβ5-L, is essential for acceleration of hydrolysis of GTP by Gαt and for normal recovery of the photoresponse.

To investigate the functional role of RGS-9 in retinal photoreceptors, we inactivated both alleles of the RGS-9 gene by replacing a 5.5-kilobase (kb) genomic fragment containing exons 2 to 4 with a neomycin resistance marker, MC1neopA (Fig. 1a; see Methods). Heterozygous and homozygous knockout mice had normal retinal morphologies up to eight months of age (data not shown). Western blot analysis showed that the retinas of RGS9^{+/+} and RGS9^{+/-} mice contained similar amounts of RGS9-1, whereas no RGS9-1 was detectable in RGS9^{-/-} retinas (Fig. 1b). RGS9^{-/-} retinas contained normal amounts of other photoreceptor proteins including recoverin, transducin, rhodopsin kinase, guanylyl cyclase E and guanylyl cyclase F (data not shown). However, the long form of the fifth member of the G-protein β-subunit (Gβ5-L)^{11,12} was absent from RGS9^{-/-} retinas (Fig. 1b), although normal amounts of Gβ5-L mRNA were still present (Fig. 1c). Gβ5 isoforms form complexes

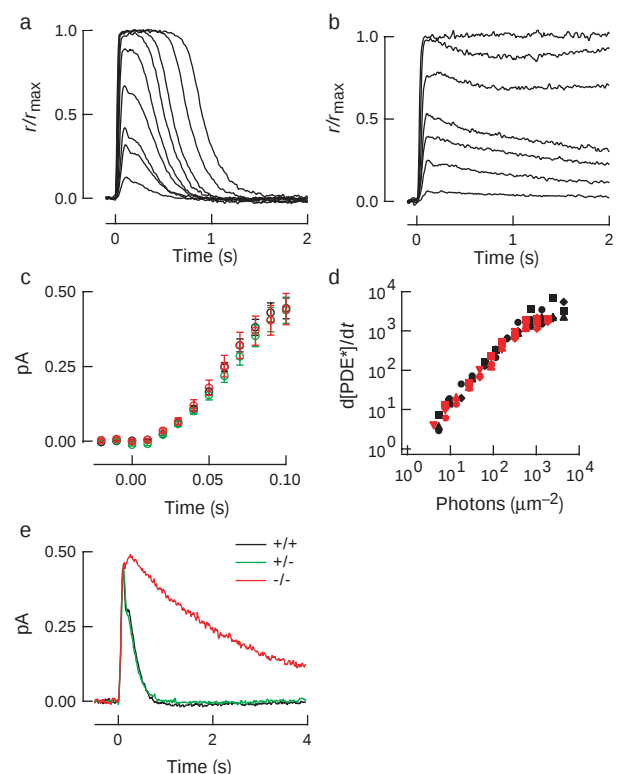


Figure 2 Slow recovery of the photoresponse in the absence of RGS9-1 and Gβ5-L. Representative families of responses from a control rod (**a**) and an RGS9^{-/-} rod (**b**) to flashes of 9.6–1350 photons μ m⁻² (**a**) or 4.0–572 photons μ m⁻² (**b**). The dark currents were 14.7 pA (**a**) and 7.7 pA (**b**). Each trace is the average of 2–31 responses. **c**, Rising phases of the population average single photon response of RGS9^{-/-} (red, $n = 16$), RGS9^{+/-} (green, $n = 29$) and control rods (black, $n = 12$). Error bars indicate s.e.m. **d**, Initial rate of change of the light-activated PDE activity (d[PDE*]/dt) in five control rods (black) and five RGS9^{-/-} rods (red) as described^{18,20,21}. **e**, Average single photon responses from **c** shown on a longer time scale. In each panel the flash was delivered at $t = 0$.

with several RGS proteins including RGS6, RGS7, RGS9 and RGS11 (refs 10, 13–15), which interact with Gβ5 through their G-protein γ-like (GGL) domains^{14–15}. The absence of Gβ5-L in RGS9^{-/-} retinas indicates that RGS9-1 may be required for the translation or the stability of Gβ5-L in photoreceptors.

In the absence of RGS9-1, hydrolysis of GTP was significantly slowed (Fig. 1d). In RGS9^{+/+} ROS, the rate constant of GTP hydrolysis (k_{inact}) by Gαt was $0.087 \pm 0.0021 \text{ s}^{-1}$ (mean $\pm$ s.e.m.); in RGS9^{-/-} ROS it was $0.026 \pm 0.0021 \text{ s}^{-1}$. The k_{inact} in RGS9^{-/-} ROS was comparable to that for isolated Gαt, indicating that RGS9-1 is indeed essential for the acceleration of hydrolysis of GTP by Gαt. Although k_{inact} for both RGS9^{+/+} and RGS9^{-/-} ROS was considerably slower than the rate of recovery of the flash responses (see below), we attribute the slow *in vitro* rate to the more than 50-fold dilution of cellular contents in the assay.

RGS9-deficient rods provided a unique opportunity to investigate the contribution of the effector γ-subunit of cyclic GMP phosphodiesterase (PDEγ) to Gαt deactivation (Fig. 1d). When the diluted ROS used in our assays were supplemented with PDEγ, the rate constant of hydrolysis of GTP in RGS9^{+/+} ROS increased to $0.148 \pm 0.011 \text{ s}^{-1}$, consistent with previous reports that PDEγ can speed up hydrolysis of GTP by Gαt^{16,17}. In contrast, exogenous PDEγ had no effect on hydrolysis of GTP in RGS9^{-/-} ROS ($k_{\text{inact}} = 0.027 \pm 0.002 \text{ s}^{-1}$), indicating that PDEγ by itself has no GTPase accelerating activity. The effect of exogenous PDEγ in control ROS membranes therefore depends on RGS9-1, consistent with previous reports that the G-protein–effector complex is the preferred target for RGS proteins^{17,18}.

To investigate the function of RGS9-1 in phototransduction, we recorded from individual rod cells using suction electrodes¹⁹. Responses from RGS9^{+/+} (Fig. 2a) and RGS9^{-/-} (Fig. 2b) rods revealed a specific defect in the recovery phase of the RGS9^{-/-} responses over a wide range of flash strengths (Fig. 2b, d). Single exponential fits to the recovery phase of the dim flash response gave time constants of about 0.2 s (+/+) and 2.5 s (-/-), and the integration time of the dim flash response was prolonged seven-fold in the knockout rods (Table 1).

The deletion of RGS9 had no effect on the amplitude of flash responses, as indicated by the values of flash sensitivity, single photon response amplitude and half-saturating flash strength listed in Table 2. The rising phases of the average single photon responses were virtually identical in +/+, +/- and -/- rods (Fig. 2c), and the initial rate of change of the light-dependent PDE activity^{18,20,21} was indistinguishable in control and knockout rods over a wide range of flash strengths (Fig. 2d). This indicates that

interactions between rhodopsin and transducin, as well as transducin and PDE, proceeded normally in RGS9^{-/-} rods.

Although the rising phases of RGS9^{-/-} and RGS9^{+/+} responses coincided, RGS9^{-/-} dim flash responses continued to rise for a longer time, so that they reached peak amplitude slightly later, on average, than control responses (Table 1), consistent with a prolonged lifetime of activated Gαt. The rate of activation in RGS9^{-/-} responses slowed about 170 ms after the flash, so that the response amplitude was not significantly larger than that of controls (Table 2). This slowing may result from a decrease in the catalytic activity of rhodopsin, as rhodopsin kinase and arrestin begin to deactivate rhodopsin by this time^{20,22}.

The effect of the RGS9 deletion resembles that of dialysing GTPγS into truncated salamander rods²³ and is consistent with the sole deficit in RGS9^{-/-} rods being a prolongation of the lifetime of activated Gαt. The RGS9^{-/-} phenotype indicates that RGS9 is essential for the high rate of GTP hydrolysis in normal rods.

In RGS9^{-/-} rods, unlike control rods, the rate of recovery of the

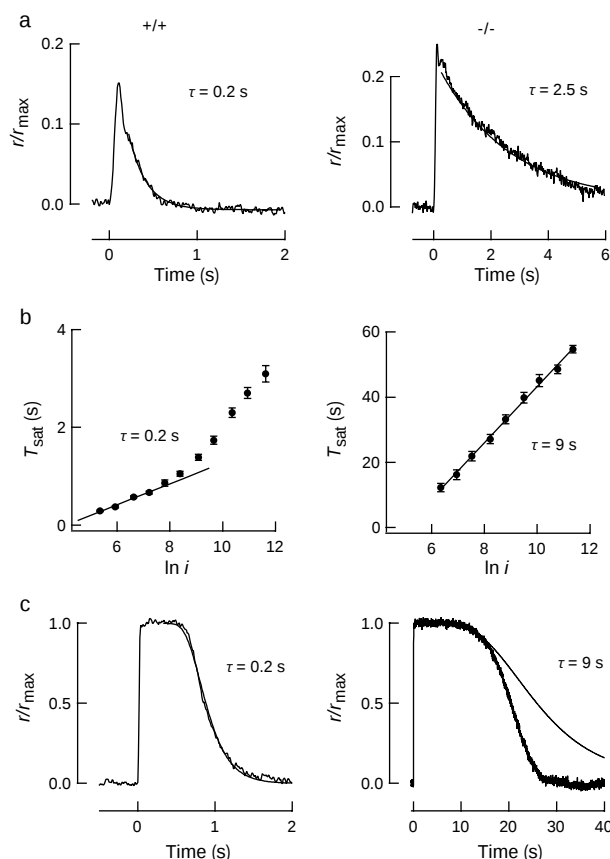


Figure 3 Dependence of recovery on flash strength. **a**, Representative mean dim flash responses from a control rod (left) and an RGS9^{-/-} rod (right). The falling phases were fitted by single exponential functions. **b**, Rate-limiting time constant of recovery from saturating flashes was greatly slowed in RGS9^{-/-} rods. In control rods, the rate-limiting time constant of recovery from saturating flashes was 0.2 s (left, collected results from 13 cells, 6–13 measurements used to determine each mean value). In RGS9^{-/-} rods (right), the recovery time constant was 8.6 s (collected results from 8 cells, 4–8 measurements used to determine each mean value). Error bars indicate s.e.m. **c**, In a control rod, the entire time course of recovery of the saturating response (left; same rod as in **a**) was fitted by a single exponential function with a 0.2-s time constant (thick line). However, in the saturating response of a RGS9^{-/-} rod (right; same rod as in **a**), a single exponential function with a 9-s time constant did not fit the entire recovery phase. Instead, the deactivation speeded as the response recovered. Flash strengths for the control responses were 9.55 (**a**) and 1,350 (**c**) photons μm^{-2} , those for RGS9^{-/-} responses were 13.6 (**a**) and 572 (**c**) photons μm^{-2} . Dark currents for the control and RGS9^{-/-} rods were 10.8 and 7.7 pA, respectively.

Table 1 Kinetic characteristics of control and RGS9^{-/-} flash responses

Strains	Integration time (ms)	Time to peak (ms)	Dim flash τ_{rec} (s)	Saturating flash τ_{rec} (s)
RGS9 ^{+/+}	267 $\pm$ 26 (12)	104 $\pm$ 3 (12)	0.18 $\pm$ 0.01 (12)	0.27 $\pm$ 0.02 (13)
RGS9 ^{+/-}	250 $\pm$ 26 (29)	123 $\pm$ 8 (30)	0.18 $\pm$ 0.02 (29)	0.22 $\pm$ 0.01 (24)
RGS9 ^{-/-}	1,881 $\pm$ 129 (15)	283 $\pm$ 39 (18)	2.60 $\pm$ 0.32 (18)	8.99 $\pm$ 0.22 (8)

All values are mean $\pm$ s.e.m. The number of rods used to determine each parameter is given in parentheses. The dim flash τ_{rec} was measured by fitting a single exponential function to the falling phase of a dim flash response. Saturating flash τ_{rec} was determined as the slope of the linear relation between the time spent in saturation and the natural log of the flash strength^{24,25}.

Table 2 Sensitivity characteristics of control and RGS9^{-/-} flash responses

Strains	i_0^* (photons μm^{-2})	Single photon response amplitude (pA)	Flash sensitivity (pA per photons per μm^{-2})	$I_{\text{d}}^\dagger$ (pA)
RGS9 ^{+/+}	66.7 $\pm$ 6.4 (12)	0.46 $\pm$ 0.04 (12)	0.16 $\pm$ 0.03 (12)	12.3 $\pm$ 0.8 (13)
RGS9 ^{+/-}	77.0 $\pm$ 3.5 (32)	0.47 $\pm$ 0.05 (29)	0.11 $\pm$ 0.02 (20)	10.5 $\pm$ 0.6 (32)
RGS9 ^{-/-}	65.0 $\pm$ 6.9 (18)	0.55 $\pm$ 0.08 (16)	0.11 $\pm$ 0.01 (16)	8.8 $\pm$ 0.5 (23)

* Flash strength that elicited a half-maximal response.

† Inward current in darkness. I_{d} was somewhat smaller in -/- rods than in +/+, which might be explained by a small reduction in the length of the outer segment. Such an effect has been observed in other rods with impairments in deactivation of the light response^{20,22,30}.

flash responses slowed progressively as the flash strength was raised. Dim flash responses of RGS9^{-/-} rods recovered exponentially with a 2.5-s time constant (Fig. 3a, right; Table 1). However, the time constant of the rate-limiting deactivation step in saturated responses, given by the slope of the relationship between the time that the response remained saturated (T_{sat}) and the log of the flash strength^{24,25}, was 9 s (Fig. 3b, right). Although saturating responses of RGS9^{-/-} rods began to decline with a time constant of 9 s, the deactivation speeded as the response recovered (Fig. 3c, right). In sharp contrast, both dim and moderately strong saturating responses of +/+ rods recovered exponentially with a time constant of 0.2 s (Fig. 3a, c, left; Table 1) and the time constant of the rate-limiting deactivation step in saturated responses was also 0.2 s (straight line in Fig. 3b, left, up to $\ln i = 8.5$). At flash strengths exceeding $\ln i = 8.5$ in control rods, the slope steepened to 0.7 s. This effect is attributable to slowed deactivation of the cascade when more than one photoisomerization occurs per disc face. Multiple photoisomerizations per disc face would cause the concentration of activated G α t to exceed that of PDE γ and thereby decrease the rate at which G α t deactivates²⁵.

A striking feature of the RGS9^{-/-} phenotype was the dependence of the recovery rate on flash strength at strengths below those that would cause multiple photoisomerizations per disc face. Indeed, in all 15 cells examined, the recovery rate appeared to vary continuously with the response amplitude, accelerating as the response fell (for example, Fig. 3c, right). In control rods at similar flash strengths recovery was faster and governed by a single time constant. By assuming that the rate of G α t deactivation varies linearly with the inward current in RGS9^{-/-} rods, we could simulate their responses over a wide range of flash strengths (Fig. 4, right), whereas the

responses of control rods were well fitted by assuming that transducin deactivated with a single time constant of 0.2 s (Fig. 4, left). Our interpretation is that RGS9^{-/-} rods may contain a relatively weak GTPase accelerating factor whose activity varies with membrane current because it depends on the concentration of calcium or cyclic GMP, both of which vary with the membrane current. The fact that the dim flash response in RGS9^{-/-} rods fell with a time constant 10-fold longer than that in control rods indicates that the residual accelerating factor might contribute little to the GTPase activity of normal rods.

We conclude that RGS9-1 is required for normal deactivation of transducin and response recovery in retinal rods. In the absence of RGS9-1, PDE γ has no GTPase accelerating activity. Thus, RGS9-1 is the essential GTPase accelerating factor in rods and the G α t–PDE γ complex is its preferred target. In addition, our results indicate that RGS9-1 is required for stable expression of G β 5-L. This is the first demonstration, to our knowledge, that an RGS protein is required for the proper function of a G-protein cascade in mammals. □

Methods

Transgenic mice

RGS9^{-/-} mice were generated by standard procedures²⁶. Transfected embryonic stem cell clones derived from 129SvJ mice were screened for homologous recombination by polymerase chain reaction (PCR) using a primer inside the MC1neopA cassette (neo-3, 5'-TCGCCGCTCCCGATTGCGAGCGCA-3') and a primer outside the 1.3-kb short arm (RGS9-ko1, 5'-GAGAAAGGATCCAGGAACCTGTAG-3'). Positive clones were micro-injected into blastocysts derived from C57/B6 to generate chimaeric mice, which were then mated with either C57/B6 or 129SvJ to produce hemizygous knockouts (RGS9^{-/-}). RGS9^{-/-} mice were produced by inbreeding of RGS9^{+/-} mice. All experimental procedures complied with NIH guidelines as approved by the Institutional Animal Care and Use Committee of the California Institute of Technology.

Immunoblots

Proteins were detected by western blotting of 15 μ g total retinal proteins. The polyclonal antibodies used for detecting photoreceptor-specific proteins were: CT-215 (anti-G β 5, used at 1:2,500 dilution); CT-317 (anti-RGS9-1, 1:1,000); T α -1A (anti-G α t, 1:2,000); 8585 (anti-rhodopsin kinase, 1:3,000); 8698 (anti-recoverin, 1:5,000); α -GC-E (anti-guanylyl cyclase E, 1:10,000); and α -GC-F (anti-guanylyl cyclase F, 1:10,000).

Single-turnover GTPase assay

Retinas were isolated from dark-adapted mice under dim red light, and ROS were isolated from 15 retinas using standard discontinuous sucrose gradient centrifugation procedures. ROS were hypotonically shocked in 90 μ l H₂O for 10 s by vortexing and then resuspended in 100 μ l GAPN buffer (10 mM Tris, pH 7.4, 100 mM NaCl, 2 mM MgCl₂, solid PMSF and 1 mM DTT). Single turnover G α t GTPase assays were carried out in triplicate as described²⁷ with or without exogenous PDE γ . Final concentrations were: Rhodopsin, 20 μ M; recombinant (His)₆-PDE γ (ref. 28), 0 or 1.33 μ M; AMP-PNP, 50 μ M; in GAPN buffer.

Electrophysiology

All mice used in physiological experiments were housed in 12-h cyclic light and dark-adapted overnight before we isolated the retinas and made recordings. Small pieces of retina were perfused with bicarbonate buffered Locke's solution (112.5 mM NaCl, 3.6 mM KCl, 2.4 mM MgCl₂, 1.2 mM CaCl₂, 10 mM HEPES (pH 7.4), 0.02 mM EDTA, 20 mM NaHCO₃, 3 mM Na₂-succinate, 0.5 mM Na-glutamate, 10 mM glucose and 0.1% vitamin and amino acids supplement solution (Sigma)) warmed to 34–37 °C and bubbled with 95% O₂/5% CO₂. We drew outer segments of single rods into a suction electrode that contained 140 mM NaCl, 3.6 mM KCl, 2.4 mM MgCl₂, 1.2 mM CaCl₂, 3 mM HEPES (pH 7.4), 0.02 mM EDTA and 10 mM glucose. Brief flashes (10 ms) of 500-nm light of calibrated intensity were used to stimulate the rods. Single-cell recording procedures were done as described¹⁸.

The time spent in saturation was taken as the interval between the midpoint of the flash and the point at which the response recovered to 10% of its initial value. We obtained the form and amplitude of the single photon response by comparing the squared mean and the time-dependent ensemble variance of a series of responses to dim flashes of fixed strength¹⁹. The mean linear response to at least 25 dim flashes was squared and scaled so that its rising phase coincided with that of the time-dependent variance. The scaling factor for the squared mean response is $1/n$ where n is the mean number of photoisomerizations per trial¹⁹. The mean response, divided by n , gave the estimated form and amplitude of the single photon response.

The time course of the light-activated PDE activity, PDE $^*(t)$, was calculated as described^{18,20,21}. The initial rate of change of PDE activation, $d[\text{PDE}^*(t)]/dt$, was determined by measuring the slope of a straight line fitted to PDE $^*(t)$ between the time of the flash and the time at which the PDE activity reached its maximum value, a time window which varied from 100 ms for dim flashes to 30 ms for saturating flashes.

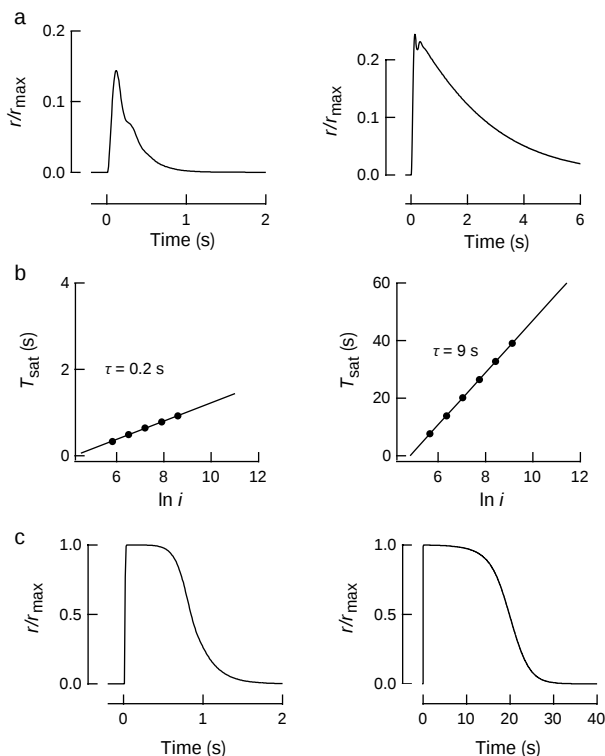


Figure 4 Simulated flash responses of rods. Left, control; right, RGS9^{-/-} rod. **a**, Dim and **c**, saturating flashes. For +/+ responses, G α t was assumed to deactivate with a single time constant of 0.2 s. For RGS9^{-/-} responses, G α t was assumed to deactivate with a time constant that increased linearly with the inward current. In all other respects, control and RGS9^{-/-} rods were assumed to be identical (see Methods). **b**, Dependence of the time in saturation (T_{sat}) on flash strength calculated for control and RGS9^{-/-} simulated responses.

Mathematical simulations

An Igor-based modelling program (written by F. Rieke, University of Washington) was adapted to compute the theoretical flash responses for control and RGS9^{-/-} responses. In this simulation, diffusion was ignored and the interior of the outer segment was assumed to be well-stirred. Theoretical flash responses were calculated using equations 4, 5, 6 and 7 in ref. 29. Rhodopsin activity was assumed to undergo first-order decay, and channel activation was assumed to vary with the cube of the cGMP concentration. For control responses, the time constant of PDE (or, equivalently G α t) deactivation was 0.2 s, whereas in RGS9^{-/-} rods the time constant was assumed to vary linearly with the inward current, being 9 s when the current was zero and 2.5 s when the current was at its dark level. All other parameters used to fit +/+ and RGS9^{-/-} were identical and were: rhodopsin lifetime: 0.09 s; basal PDE activity: 9 s⁻¹; dark [Ca²⁺]: 400 nM; minimum [Ca²⁺] in saturating light: 50 nM; Hill coefficient of guanylate cyclase for Ca²⁺: 2.3; half-maximal activation of guanylate cyclase by Ca²⁺: 400 nM; Hill coefficient of channels for cGMP: 3; time constant of Na⁺/Ca²⁺, K⁺ exchange: 40 ms. To simulate the responses of the cells in Fig. 3, the dark currents were taken as 10.8 pA (+/+) and 7.7 pA (RGS9^{-/-}).

Received 5 October; accepted 24 November 1999.

- Arshavsky, V. Y. & Pugh, E. N. Lifetime regulation of G protein-effector complex: emerging importance of RGS proteins. *Neuron* **20**, 11–14 (1998).
- Berman, D. M. & Gilman, A. G. Mammalian RGS proteins: barbarians at the gate. *J. Biol. Chem.* **273**, 1269–1272 (1998).
- Wieland, T. & Chen, C.-K. Regulator of G-protein signaling: a novel protein family involved in timely deactivation and desensitization of signaling via heterotrimeric G-proteins. *Naunyn-Schmiedeberg's Arch. Pharmacol.* **360**, 14–20 (1999).
- De Vries, L. & Farquhar, M. G. RGS proteins: more than just GAPs for heterotrimeric G proteins. *Trends Cell Biol.* **9**, 138–144 (1999).
- Nekrasova, E. R. et al. Activation of transducin guanosine triphosphatase by two proteins of the RGS family. *Biochemistry* **36**, 7638–7643 (1997).
- Chen, C.-K., Wieland, T. & Simon, M. I. RGS-r, a retinal specific RGS protein, binds an intermediate conformation of transducin and enhances recycling. *Proc. Natl Acad. Sci. USA* **93**, 12885–12889 (1996).
- Fauverbert, E. & Hurley, J. B. The core domain of a new retina specific RGS protein stimulates the GTPase activity of transducin in vitro. *Proc. Natl Acad. Sci. USA* **94**, 2945–2950 (1997).
- He, W., Cowan, C. W. & Wensel, T. G. RGS9, a GTPase accelerator for phototransduction. *Neuron* **20**, 95–102 (1998).
- Cowan, C. W., Fariss, R. N., Sokal, I., Palczewski, K. & Wensel, T. G. High expression levels in cones of RGS9, the predominant GTPase accelerating protein of rods. *Proc. Natl Acad. Sci. USA* **95**, 5351–5356 (1998).
- Makino, E. R., Handy, J. W., Li, T. & Arshavsky, V. Y. The GTPase activating factor for transducin in rod photoreceptors is the complex between RGS9 and type 5 G protein β subunit. *Proc. Natl Acad. Sci. USA* **96**, 1947–1952 (1999).
- Watson, A. J., Katz, A. & Simon, M. I. A fifth member of the mammalian G-protein β -subunit family. Expression in brain and activation of the β 2 isotype of phospholipase C. *J. Biol. Chem.* **269**, 22150–22156 (1994).
- Watson, A. J., Aragay, A. M., Slepak, V. Z. & Simon, M. I. A novel form of the G protein β subunit G β 5 is specifically expressed in the vertebrate retina. *J. Biol. Chem.* **271**, 28154–28160 (1996).
- Cabrera, J. L., de Freitas, F., Satpaev, D. K. & Slepak, V. Z. Identification of the G β 5-RGS7 protein complex in the retina. *Biochem. Biophys. Res. Commun.* **249**, 898–902 (1998).
- Snow, B. E. et al. A G protein γ subunit-like domain shared between RGS11 and other RGS proteins specifies binding to G β 5 subunits. *Proc. Natl Acad. Sci. USA* **95**, 13307–13312 (1998).
- Levay, K., Cabrera, J. L., Satpaev, D. K. & Slepak, V. Z. G β 5 prevents the RGS7-G α o interaction through binding to a distinct G γ -like domain found in RGS7 and other RGS proteins. *Proc. Natl Acad. Sci. USA* **96**, 2503–2507 (1999).
- Arshavsky, V. Y. & Bownds, M. D. Regulation of deactivation of photoreceptor G protein by its target enzyme and cGMP. *Nature* **357**, 416–417 (1992).
- Pages, F., Deterre, P. & Pfister, C. Enhanced GTPase activity of transducing when bound to cGMP phosphodiesterase in bovine retinal rods. *J. Biol. Chem.* **267**, 22018–22021 (1992).
- Tsang, S. H. et al. Role for the target enzyme in deactivation of photoreceptor G protein in vivo. *Science* **282**, 117–121 (1998).
- Baylор, D. A., Lamb, T. D. & Yau, K. W. Responses of retinal rods to single photons. *J. Physiol. (Lond.)* **288**, 613–634 (1979).
- Chen, C.-K. et al. Abnormal photoresponses and light-induced apoptosis in rods lacking rhodopsin kinase. *Proc. Natl Acad. Sci. USA* **96**, 3718–3722 (1999).
- Lamb, T. D. & Pugh, E. N. A quantitative account of the activation steps involved in phototransduction in amphibian photoreceptors. *J. Physiol.* **449**, 719–758 (1992).
- Xu, J. et al. Prolonged photoresponses in transgenic mouse rods lacking arrestin. *Nature* **389**, 505–509 (1997).
- Sagoo, M. S. & Lagnado, L. G-protein deactivation is rate-limiting for shut-off of the phototransduction cascade. *Nature* **389**, 392–395 (1997).
- Pepperberg, D. R. et al. Light-dependent delay in the falling phase of the retinal rod photoresponse. *Vis. Neurosci.* **8**, 9–18 (1992).
- Lyubarsky, A. L. & Pugh, E. N. Recovery phase of the murine rod photoresponse reconstructed from electrophoretographic recordings. *J. Neurosci.* **16**, 563–571 (1996).
- Ramirez-Solis, R., Davis, A. C. & Bradley, A. in *Methods in Enzymology: Guide to Techniques in Mouse Development* (eds Wassarman, P. M. & DePamphilis, M. L.) 855–878 (Academic, San Diego, 1993).
- Cowan, C. W., Wensel, T. G. & Arshavsky, V. Y. in *Methods in Enzymology: Vertebrate Phototransduction and the Visual Cycle* (ed. Palczewski, K.) 524–538 (Academic, San Diego, 2000).
- Wieland, T., Chen, C.-K. & Simon, M. I. The retinal specific protein RGS-r competes with the γ subunit of cGMP phosphodiesterase for the α subunit of transducin and facilitates signal termination. *J. Biol. Chem.* **272**, 8853–8856 (1997).
- Rieke, F. & Baylor, D. A. Single-photon detection by rod cells of the retina. *Rev. Mod. Phys.* **70**, 1027–1036 (1998).

30. Makino, C. L., Flannery, J. G., Chen, J. & Dodd, R. L. in *Photostasis and Related Phenomenon* (eds Williams, T. P. & Thistle, A. B.) 129–151 (Plenum, New York, 1998).

Acknowledgements

We thank members of the Caltech Transgenic Core Facility for their technical support, and R. Lefkowitz, J. Hurley and H. Jurgen-Fuller for antibodies. This work was supported by grants from the NIH.

Correspondence and requests for materials should be addressed to M.I.S. (e-mail: simonm@starbase1.caltech.edu).

Food and metabolic signalling defects in a *Caenorhabditis elegans* serotonin-synthesis mutant

Ji Ying Sze^{*†}, Martin Victor[‡], Curtis Loer[§], Yang Shi[‡] & Gary Ruvkun^{*}

^{*} Department of Molecular Biology, Massachusetts General Hospital, Department of Genetics, Harvard Medical School, Boston, Massachusetts 02114, USA

[‡] Department of Pathology, Harvard Medical School, Boston, Massachusetts 02115, USA

[§] Department of Biology, University of San Diego, San Diego, California 92110, USA

The functions of serotonin have been assigned through serotonin-receptor-specific drugs and mutants^{1,2}; however, because a constellation of receptors remains when a single receptor subtype is inhibited, the coordinate responses to modulation of serotonin levels may be missed. Here we report the analysis of behavioural and neuroendocrine defects caused by a complete lack of serotonin signalling. Analysis of the *C. elegans* genome sequence showed that there is a single tryptophan hydroxylase gene (*tph-1*)—the key enzyme for serotonin biosynthesis. Animals bearing a *tph-1* deletion mutation do not synthesize serotonin but are fully viable. The *tph-1* mutant shows abnormalities in behaviour and metabolism that are normally coupled with the sensation and ingestion of food: rates of feeding and egg laying are decreased; large amounts of fat are stored; reproductive lifespan is increased; and some animals arrest at the metabolically inactive dauer stage. This metabolic dysregulation is, in part, due to downregulation of transforming growth factor- β and insulin-like neuroendocrine signals. The action of the *C. elegans* serotonergic system in metabolic control is similar to mammalian serotonergic input to metabolism and obesity².

Serotonin is synthesized from tryptophan by an enzymatic pathway: tryptophan hydroxylase (TPH) catalyses the rate-limiting first step, and the dual functional 5-hydroxytryptophan (5HTP)/L-dopa decarboxylase then matures 5HTP to serotonin (Fig. 1a)¹. There is one probable orthologue of mammalian tryptophan hydroxylase in the *C. elegans* genome, ZK1290.2, which we call *tph-1* (Fig. 1b). Two other *C. elegans* aromatic amino-acid hydroxylase (AAAH) gene family members are B0432.5, a probable tyrosine hydroxylase that catalyses dopamine synthesis and is expressed in dopaminergic neurons (R. Lints and S. Emmons, personal communication), and K08F8.4, a probable phenylalanine hydroxylase that is not expressed in neurons³. *tph-1* shares higher amino-acid identity in the 330-amino-acid catalytic domain with mammalian tryptophan hydroxylase (61%) than with mammalian phenylalanine (53%) or

[†] Current address: Department of Anatomy and Neurobiology, College of Medicine, University of California, Irvine, California 92697-1280, USA.

tyrosine (51%) hydroxylases (Fig. 1c)⁴. Short protein motifs are also conserved between *tph-1* and mammalian tryptophan hydroxylases but not the other hydroxylases (Fig. 1c); these regions may mediate the enzyme specificity for tryptophan as opposed to other aromatic amino acids.

A green fluorescent protein gene fusion to *tph-1* (*tph-1::GFP*) is expressed in *C. elegans* serotonergic neurons⁵: NSM, ADF, hermaphrodite-specific HSN, male-specific CP, and, rarely, AIM and RIH (Fig. 2a,b,c). *tph-1::GFP* expression in most of these neurons is observed by the L1 stage but expression in the late-maturing HSN and CP neurons is observed in the adult stage. Thus, the expression pattern of *tph-1::GFP* is consistent with its biochemical function in the synthesis of serotonin.

We generated a *tph-1* deletion mutation by sib-screening mutagenized animals with polymerase chain reaction (PCR) primers from the *tph-1* gene⁶. The *tph-1(mg280)* mutant allele is a 1,305-base-pair (bp) deletion that removes conserved amino acids in both the regulatory and catalytic domains, as defined by the analysis of mammalian tryptophan hydroxylase (Fig. 1b)⁴. In addition, the predicted splicing of exon 1 to exon 8 from the extant *tph-1(mg280)* mutant transcript generates a frameshift that terminates translation 16 codons after the exon-8 splice acceptor site, upstream of more conserved catalytic domain sequences. Therefore, on the basis of the conserved exons deleted from the gene and the additional exons predicted to be deleted by the aberrant splicing, *tph-1(mg280)* is a probable null mutation.

tph-1(mg280) homozygous mutant animals are viable, but show a variety of behavioural and metabolic defects (see below). Consistent with the molecular prediction that the *tph-1(mg280)* mutation eliminates the gene activity, mutant animals accumulate no detectable serotonin as determined by anti-serotonin immunofluorescence (Fig. 2d). A 5.8-kilobase (kb) genomic DNA fragment that contains *tph-1* regulatory and coding sequences (Fig. 1b) restores serotonin production to *tph-1(mg280)* mutant animals (Fig. 2e). The lack of detectable serotonin in *tph-1(mg280)* mutant animals suggests that even though all three *C. elegans* AAAH genes are closely related, they are not functionally redundant, and that *tph-1* is the main and perhaps the only biosynthetic pathway for the production of serotonin. In support of no crosstalk between tryptophan hydroxylase and tyrosine hydroxylase enzymes, *tph-1(mg280)* mutant animals have a normal pattern of dopamine distribution (Fig. 2g), and the pattern of serotonin accumulation is not changed in the tyrosine hydroxylase deficient mutant *cat-2* (R. Lints and S. Emmons, personal communication; J.Y.S. and G.R., unpublished data). Serotonin signalling is not necessary for the gross pathfinding or survival of the serotonergic neurons. As revealed by the *tph-1::GFP* fusion gene, the serotonergic neurons have normal morphology and pathfinding in *tph-1(mg280)* mutant animals (Fig. 2f).

Exogenous serotonin modulates several *C. elegans* behaviours in a manner similar to changing levels of the food^{7,8}. Previous reports have shown that mutants with decreased serotonin and dopamine levels display egg-laying and pharyngeal pumping defects⁸⁻¹². *tph-1(mg280)* animals display several behavioural defects that are associated with starvation: on a bacterial lawn, the mutant pumps slower and retains more eggs than wild type animals (Fig. 3). A *tph-1(+)* transgene rescues these *tph-1(mg280)* behavioural defects (Fig. 3a,b). The depression of pharyngeal pumping and egg laying is similar to that caused by starvation of wild-type animals^{8,10}. In the absence of food, the pharyngeal pumping rate of *tph-1(mg280)* is similar to that of the wild type (Fig. 3a), showing that serotonin is necessary for the upregulation of the pharyngeal pumping in the presence of food but not for basal activity. Increased *tph-1* gene dosage causes increased egg laying compared with wild-type animals (Fig. 3b).

Food level not only modulates *C. elegans* motor outputs, it also affects a metabolism-regulating neurosecretory axis. Food is one of

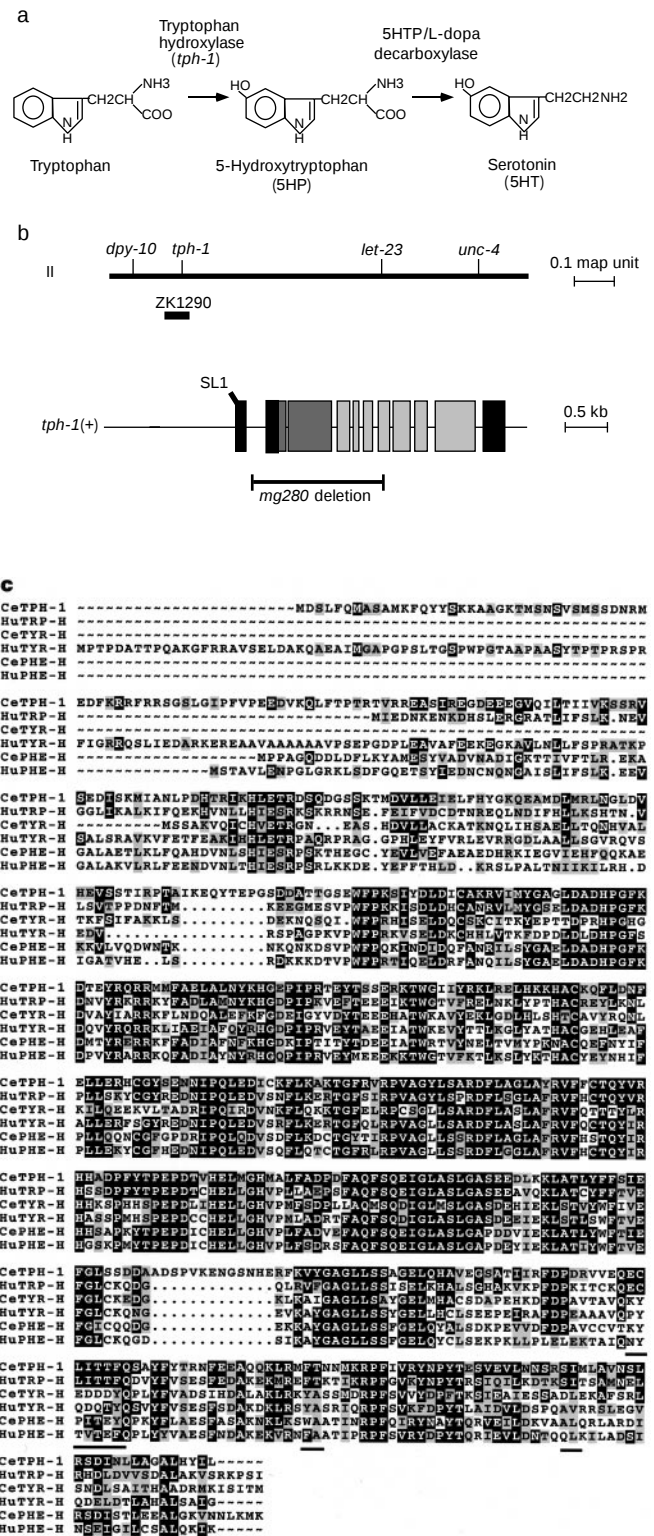


Figure 1 *tph-1* encodes a tryptophan hydroxylase. **a**, The serotonin biosynthetic pathway¹. **b**, A 5.8-kb genomic fragment from 3.1-kb upstream to 90-bp downstream of the *tph-1* transcript rescues *tph-1(mg280)*. The coding regions are in boxes and the non-coding regions are shown as lines. The regulatory and catalytic domains are highlighted by light and dark hatching, respectively⁴. The *mg280* deletion is depicted. **c**, A comparison of *C. elegans* TPH-1 (AF135186) with human tryptophan hydroxylase (HuTRP-H, P17752), *C. elegans* probable tyrosine (CeTYR-H, B0432.5) and phenylalanine (CePHE-H, K08F8.4) hydroxylases, human tyrosine (HuTYR-H, P07101) and phenylalanine (HuPHE-H, P00439) hydroxylases aligned using PILEUP. Motifs conserved in the tryptophan hydroxylases are underlined.

the sensory inputs (a secreted pheromone and growth temperature are others) that control whether wild-type *C. elegans* enter the metabolically shifted, non-feeding, non-reproducing dauer stage and store large amounts of fat^{13,14}. We found that 10–15% of *tph-1(mg280)* animals arrest at the dauer stage even in the presence of abundant food, unlike wild-type animals (Fig. 3c). Another 10–15% of *tph-1(mg280)* animals form partial dauer larvae. The dauer arrest phenotype is suppressed by growing *tph-1(mg280)* animals on plates supplied with serotonin (Fig. 3c). The dauer arrest phenotype is independent of temperature, suggesting that temperature modulation of dauer arrest may in part be serotonergic¹³. Although only a minority of the *tph-1(mg280)* animals arrest at the dauer stage, 92% of L2 and L3 stage animals ($n=109$) accumulate larger stores of fat than wild-type animals ($n=111$) (Fig. 4a,b). The fat storage phenotype is rescued in *tph-1(mg280)* animals carrying a *tph-1(+)* transgene (68% of *tph-1(mg280);Ex tph-1(+)* transgenic animals have wild-type fat stores ($n=53$)).

Insulin-like and DAF-7 transforming growth factor (TGF)- β signalling pathways act in parallel to regulate *C. elegans* metabolism and development; disruption of either pathway is sufficient to cause constitutive arrest at the dauer stage¹⁵ and a switch to fat storage

metabolism¹⁴. Enhancement and suppression genetics as well as molecular experiments implicate serotonin inputs to both the TGF- β and insulin-like pathways. A *daf-7::GFP* fusion gene is predominantly expressed in the ASI sensory neurons in well-fed animals, and the level of expression is downregulated by dauer pheromone and upregulated by food signals^{15,16}. There is a serotonergic input to the DAF-7 pathway at the level of production of the DAF-7 TGF- β neuroendocrine signal. First, using an integrated *daf-7::GFP* fusion gene, 57% ($n=79$) of *tph-1(mg280)* animals do not express detectable *daf-7::GFP*, compared with 9% ($n=94$) in wild-type. Second, of the animals that express *daf-7::GFP*, the expression level is decreased 6.5-fold in *tph-1(mg280)* animals ($n=19$) (Fig. 4d,e). Consistent with the upregulation of DAF-7 expression by serotonin, *tph-1(mg280)* strongly enhances dauer arrest of *daf-7(e1372)* at low temperature (Fig. 3d), and exogenous serotonin reverses the *tph-1* enhancement of *daf-7* (in serotonin 78% of *tph-1(mg280);daf-7(e1372)* animals ($n=102$) recover from dauer versus 23% in the absence of serotonin ($n=58$)). Conversely, high *tph-1* gene dosage potently interferes with dauer arrest induced by *daf-7(e1372)* at the non-permissive temperature (Fig. 3d). Thus, serotonin signalling is not only necessary for reproductive growth in the absence of *daf-7* signalling at low temperature, it is also sufficient to partially bypass the need for *daf-7* signalling at high temperature. A simple explanation for the suppression of *daf-7(e1372)* dauer arrest by high *tph-1* gene dosage is that the increased serotonin levels cause increased insulin-like signalling (see below) which partially compensates for a lack of DAF-7 TGF- β signalling. Serotonin levels may normally increase at low temperature to signal reproductive development and decrease at high temperature to enhance dauer arrest. There

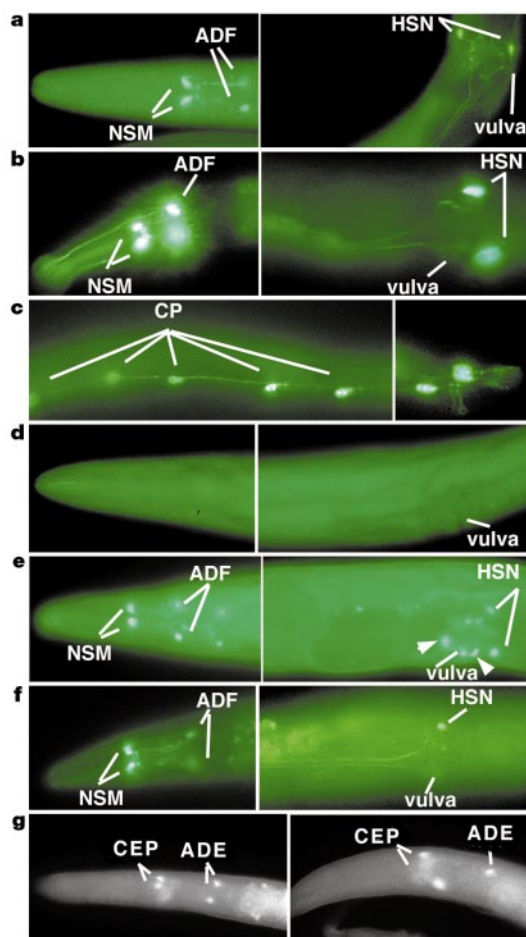


Figure 2 *tph-1* is expressed in serotonergic neurons. **a**, Anti-serotonin antibody staining in a wild-type hermaphrodite. **b**, Expression of an integrated *tph-1::GFP* fusion gene in a wild-type hermaphrodite. **c**, Expression of *tph-1::GFP* in a wild-type male. **d**, No anti-serotonin immunofluorescence in a *tph-1(mg280)* mutant hermaphrodite. **e**, *tph-1(mg280)* animals carrying the wild-type *tph-1* transgene (*tph-1(+)*) restore anti-serotonin immunofluorescence, which is noticeably higher in a subset of the transgenic animals than in wild-type animals. **f**, The pattern of serotonergic neural projections revealed by *tph-1::GFP* is not affected by *tph-1(mg280)*. **g**, *tph-1(mg280)* (right panel) and wild-type (left panel) show similar dopamine levels as visualized by formaldehyde-induced fluorescence staining⁹.

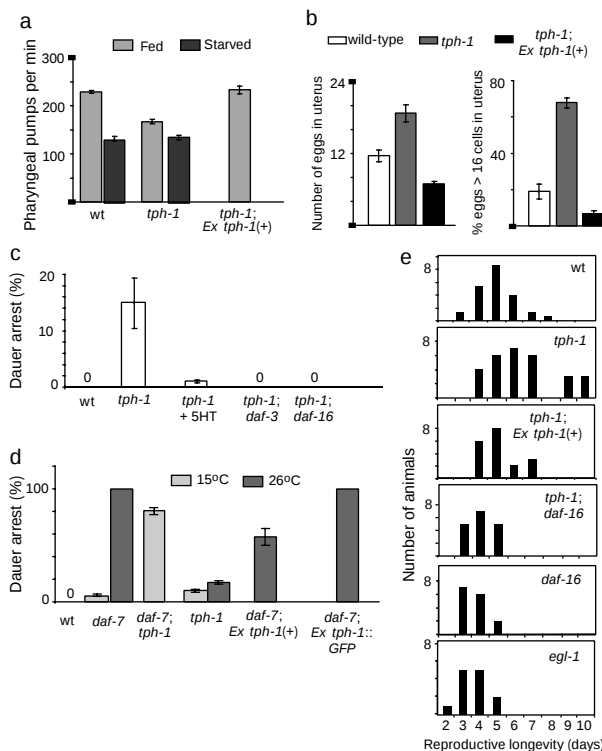


Figure 3 *tph-1(mg280)* affects food-modulated behaviours, metabolism and reproductive longevity. **a**, Decreased pharyngeal pumping in food but normal pumping without food in *tph-1(mg280)*. **b**, *tph-1(mg280)* animals carry more eggs in the uterus (left) which are released at later stages (right). *tph-1(mg280);Ex tph-1(+)* animals are rescued for these behaviours. **c**, *tph-1(mg280)* dauer arrest that is rescued by 5mM serotonin, and suppressed by mutations in *daf-3* or *daf-16*. **d**, *tph-1(mg280)* enhances *daf-7(e1372)* dauer arrest at 15°C, whereas high *tph-1* gene dosage suppresses *daf-7(e1372)* dauer arrest at 25°C. A *tph-1::GFP* fusion gene does not suppress *daf-7(e1372)* dauer arrest. **e**, *tph-1(mg280)* hermaphrodites have an extended reproductive lifespan.

are multiple probable *C. elegans* serotonin receptors¹⁷; particular receptors on the ASI neurons may be coupled to DAF-7 expression. *C. elegans* G α proteins have been implicated in control of dauer arrest and act in the *daf-7* pathway¹⁸; these G proteins may act downstream of the serotonin receptors.

There is serotonergic input to the parallel insulin-like signalling pathway and the DAF-7 TGF- β pathway. Mutations in either *daf-16*, which bypasses the need for *daf-2* insulin-like signalling¹⁹, or *daf-3*, which bypasses the need for DAF-7 TGF- β signalling²⁰, suppress the dauer arrest and fat accumulation phenotypes of *tph-1(mg280)* (Figs 3c and 4c; and data not shown). A partial decline in both DAF-7 TGF- β and insulin-like signals synergistically induces a switch to fat storage metabolism and dauer arrest¹⁹. In the *tph-1* mutant, DAF-7 TGF- β and insulin-like signals may both be decreased so that bypass of either pathway by mutation in *daf-3* or *daf-16* may allow sufficient neuroendocrine signalling to relieve the normal requirement for serotonergic input. Further evidence for serotonergic input to the insulin-like pathway comes from an analysis of *tph-1(mg280)* reproductive longevity. Like *daf-2* insulin-receptor mutants^{21,22}, *tph-1(mg280)* hermaphrodites have an extended reproductive lifespan that is dependent on *daf-16* gene activity (Fig. 3e). The extended lifespan of *tph-1(mg280)* is rescued by a *tph-1(+)* transgene, showing that it is the decrease in serotonin levels that causes the increase in reproductive longevity, rather than a linked mutation. The extended reproductive lifespan is not caused by the the egg-laying defects of the *tph-1(mg280)* mutation: *daf-16* does not suppress the egg-laying defect of *tph-1(mg280)*, and an *egl-1(n986)* mutation that causes HSN programmed cell death and similar defects in egg laying²³ does not extend reproductive age (Fig. 3e). There are *C. elegans* insulin-related genes²⁴ that are expressed in neurons and other cells (S. Pierce and G. R., unpublished data). The expression or release of any of these insulins may be responsive to serotonin. In support of serotonin input to insulin signalling, serotonin enhances insulin synthesis, release²⁵ and target tissue sensitivity in mammals²⁶.

The dauer arrest and metabolic shift of *tph-1(mg280)* is probably not the result of the slower pharyngeal pumping and caloric restriction: thirteen different *eat* mutants that have slower pumping rates do not arrest at the dauer stage or synergize with *daf* mutants for dauer arrest^{27,28}. In addition, the longevity phenotypes of *eat* mutants is not suppressed by *daf-16* (ref. 28), showing that these mutants do not act in the insulin-like pathway. Finally, the decrease in *daf-7::GFP* expression in the *tph-1* mutant is evidence of a coupling between serotonin levels and neuroendocrine signals in metabolic control.

Bacterial food and low temperature may normally upregulate the production, release or response to serotonin to in turn upregulate

DAF-7 and insulin-like neuroendocrine signals for reproductive development and low fat storage. In the *tph-1* mutant, these neuroendocrine signals may be decoupled from such food and temperature inputs. Serotonin signalling has been implicated in the control of mammalian feeding and metabolism². A mouse knockout of the 5HT_{2C} subtype receptor gene causes hyperphagia and weight gain, that eventually leads to type 2 diabetes²⁹. Conversely, addition of receptor-specific serotonin agonists, or serotonin reuptake inhibitors such as fluoxetine, reduces appetite and promotes weight loss². Serotonin signalling in mammals has also been implicated in body-temperature regulation by the hypothalamus². As body-temperature regulation is tied to metabolic rate, this suggests a coupling of serotonin to neuroendocrine outputs such as insulin, as we find in *C. elegans*. Thus mammalian 5HT levels may be regulated by nutrition and other food sensory cues to in turn regulate feeding behaviours, as well as neuroendocrine outputs analogous or homologous to *C. elegans* DAF-7 TGF- β and insulin-like hormones.

The unifying feature of many of the *C. elegans* serotonergic responses is the coupling of food sensation to various motor and endocrine outputs. Food is a major reward of many animal behaviours, both innate and learned. The ancient coupling of serotonin levels to food suggests a deep mechanistic connection between the role of satiety as a reward in mammalian associative learning and the finding that serotonin has a role in modification of synaptic signalling in the invertebrate Aplysia³⁰. □

Methods

Isolation and characterization of *tph-1(mg280)*

A DNA library representing ~600,000 haploid genomes was constructed from the progeny of wild-type animals mutagenized with diepoxyoctane. Pools of library DNA were subjected to nested PCR with the primers from the *tph-1* gene to identify a deletion mutation⁶. The mutant was back-crossed with wild-type animals four times before analysis. The *tph-1* coding sequence was confirmed by sequencing of PCR-amplified complementary DNAs. *tph-1* message is *trans*-spliced with a SL1 leader sequence 25-bp upstream of the first methionine by 5' rapid amplification of cDNA ends analysis. For the rescue experiment, a 5.8-kb genomic DNA fragment encompassing 3.1 kb of 5' flanking sequence to 300 bp 3' of the *tph-1* gene (*tph-1(+)*, Fig. 1b) was PCR amplified, cloned, and injected at 75 ng μ l⁻¹ with *rol-6* pRF4 at 100 ng μ l⁻¹.

GFP expression and immunoanalysis

A DNA fragment encompassing the 3.1 kb *tph-1* 5' regulatory sequence and introns and exons to the beginning of exon 4 was PCR amplified using primers containing restriction sites, ligated to the GFP cassette pPD95.75 (A. Fire), and injected at 50 ng μ l⁻¹, with pRF4 at 100 ng μ l⁻¹. The transgene was integrated by UV irradiation and backcrossed four times to wild-type. A *daf-7::GFP* fusion gene integrated onto a chromosome (D. Riddle) was crossed into *tph-1(mg280)*. The intensity of *daf-7::GFP* was quantified using Softwares Openlab II and Photoshop. The staining with anti-serotonin antibody was done as described¹⁰. Anti-*mec-7* antibodies (M. Hamelin) controlled for fixation procedures.

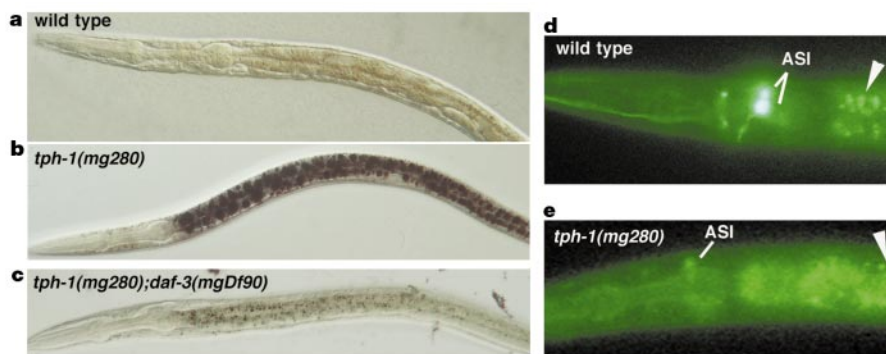


Figure 4 Metabolic defects in *tph-1(mg280)* animals. **a**, Low accumulation of fat in a wild-type L3 animal. **b**, *tph-1(mg280)* L3 animals accumulate more fat. **c**, *daf-3(mgDf90)* suppresses *tph-1(mg280)* excess fat storage at the L3 stage. **d,e**, *tph-1(mg280)* downregulates *daf-7* TGF- β expression. *daf-7::GFP* is strongly expressed in the sensory

neuron ASI in wild-type^{14,15} (**d**), but expression of the *daf-7::GFP* is reduced at least sixfold in *tph-1(mg280)* (**e**). Arrow heads show the gut autofluorescence. L1 animals are shown. Anterior is on the left.

Feeding, egg-laying, dauer arrest and reproductive lifespan assays

The rate of pharyngeal pumping at 20 °C was scored by counting pharynx terminal bulb contractions¹². The age of eggs in the uterus was estimated by counting embryonic cells in each egg. For dauer arrest, eggs laid overnight at 20 °C were shifted to 15 °C, 20 °C, or 26 °C, and scored for dauer arrest at 48 hours (26 °C), 72 hours (20 °C) and 110 hours (15 °C). The number of animals scored represents two trials of each genotype and three independent experiments. Visualized by DIC optics, the *tph-1(mg280)* dauers had dauer alae (17/18, full dauer alae; and 1/18, partial alae) and constricted pharynx (17/18), although some of them pumped occasionally in a 2 min interval (3/18). Both *daf-7(e1372);Ex tph-1(+)* and *daf-7(e1372);Ex tph-1::GFP* animals do not form dauers at 15 °C. Number of animals scored in dauer arrest assays at 20 °C: wild type, 5,442; *tph-1(mg280)*, 2,362; *tph-1(mg280)* with serotonin, 2,514; *tph-1(mg280);daf-16(mgDf50)*, 3,537; *tph-1(mg280);daf-3(mgDf90)*, 2,423. At 15 °C: wild type, 1,059; *daf-7(e1372)*, 1,743; *tph-1(mg280)*, 3,099; *daf-7(1372);tph-1(mg280)*, 1,514. At 26 °C: wild type, 1,797; *daf-7(1372)*, 901; *tph-1(mg280)*, 3,462; *daf-7(e1372);Ex tph-1(+)*, 863; *daf-7(e1372);Ex tph-1::GFP*, 1,601. The error bars represent s.e.m. of the scores from three independent experiments. For reproductive lifespan assays, animals at 20 °C were transferred every day after the L4 stage to a fresh plate seeded with bacteria and progeny during that 24-hour period were counted. Most animals lived days after they ceased reproduction. Many *tph-1(mg280)* animals died as the consequence of progeny hatching internally (12/29). Thus, their reproductive lifespans may be underestimated. Average progeny laid per animal in reproductive lifespan experiment: wild type, 287 (*n*=22); *tph-1(mg280)*, 215 (*n*=29); *tph-1(mg280);Ex tph-1(+)*, 296 (*n*=10); *daf-16(mgDf50);tph-1(mg280)*, 199 (*n*=17); *daf-16(mgDf50)*, 226 (*n*=15); *egl-1(n886)*, 144 (*n*=18). The data are the summary of 2–4 independent sets of experiments. For fat staining, animals grown at 20 °C were stained with Sudan Black as previously described¹⁴. All behavioural and metabolism analyses used animals fed *E. coli* OP50.

Received 5 November; accepted December 1999.

- Sanders-Bush, E. & Mayer, S. E. In *Goodman & Gilman's Pharmacological Basis of Therapeutics* 9th edn (eds Hardman, J. G. et al.) 249–263 (McGraw-Hill, New York, 1996).
- Leibowitz, S. F. & Alexander, J. T. Hypothalamic serotonin in control of eating behaviour, meal size, and body weight. *Biol. Psychiatry* **44**, 851–864 (1998).
- Loer, C. M., Davidson, B. D. & McKerrow, J. M. A phenylalanine hydroxylase gene from the nematode *C. elegans* is expressed in the hypodermis. *J. Neurogenet.* **13**, 157–180 (1999).
- D'Sa, C. M., Arthur, R. E. Jr & Kuhn, D. M. Expression and deletion mutagenesis of tryptophan hydroxylase fusion proteins: delineation of the enzyme catalytic core. *Neurochem.* **67**, 917–926 (1996).
- Rand, J. B. & Nonet, M. L. *Synaptic Transmission in C. elegans II* (eds Riddle, D. L. et al.) 611–643 (Cold Spring Harbor Laboratory Press, New York, 1997).
- Jansen, G., Hazendonk, E., Thijssen, K. L. & Plasterk, R. H. Reverse genetics by chemical mutagenesis in *C. elegans*. *Nature Genet.* **17**, 119–121 (1997).
- Horvitz, H. R., Chalfie, M., Trent, C., Sulston, J. E. & Evans, P. D. Serotonin and octopamine in the nematode *C. elegans*. *Science* **216**, 1012–1014 (1982).
- Avery, L. & Horvitz, H. R. Effects of starvation and neuroactive drugs on feeding in *C. elegans*. *J. Exp. Zool.* **253**, 263–270 (1990).
- Sulston, J., Dew, M. & Brenner, S. Dopaminergic neurons in the nematode *Caenorhabditis elegans*. *J. Comp. Neuro.* **163**, 215–226 (1975).
- Weinschenker, D., Garriga, G. & Thomas, J. H. Genetic and pharmacological analysis of neurotransmitters controlling egg laying in *C. elegans*. *J. Neurosci.* **15**, 6975–6985 (1995).
- Waggoner, L. E., Zhou, G. T., Schafer, R. W. & Schafer, W. R. Control of alternative behavioural states by serotonin in *Caenorhabditis elegans*. *Neuron* **21**, 203–214 (1998).
- Duerr, J. S. et al. The *cat-1* gene of *Caenorhabditis elegans* encodes a vesicular monoamine transporter required for specific monoamine-dependent behaviours. *J. Neurosci.* **19**, 72–84 (1999).
- Riddle, D. L. *Genetic and Environmental Regulation of Dauer Larva Development in C. elegans II*. (eds Riddle, D. L. et al.) 739–768 (Cold Spring Harbor Laboratory Press, New York, 1997).
- Kimura, K. D., Tissenbaum, H. A., Liu, Y. & Ruvkun, G. *daf-2*, an insulin receptor-like gene that regulates longevity and diapause in *C. elegans*. *Science* **277**, 942–946 (1997).
- Ren, P. F. et al. Control of *C. elegans* larval development by neuronal expression of a TGF- β homolog. *Science* **274**, 1389–1391 (1996).
- Schackwitz, W. S., Inoue, T. & Thomas, J. H. Chemosensory neurons function in parallel to mediate a pheromone response in *C. elegans*. *Neuron* **17**, 719–728 (1996).
- Bargmann, C. Neurobiology of the *C. elegans* genome. *Science* **282**, 2028–2033 (1998).
- Zwaal, R. R., Mendel, J. E., Sternberg, P. W. & Plasterk, R. H. A. Two neuronal G proteins are involved in chemosensation of the *C. elegans* dauer-inducing pheromone. *Genetics* **145**, 715–727 (1997).
- Ogg, S. et al. The Fork head transcription factor DAF-16 transduces insulin-like metabolic and longevity signals in *C. elegans*. *Nature* **389**, 994–999 (1997).
- Patterson, G. I., Kowek, A., Wong, A., Liu, Y. & Ruvkun, G. The DAF-3 Smad protein antagonizes TGF- β -related receptor signaling in the *C. elegans* dauer pathway. *Genes Dev.* **11**, 2679–2690 (1997).
- Kenyon, C., Chang, J., Gensch, E., Rudner, A. & Tabtiang, R. A. *C. elegans* mutant that lives twice as long as wild type. *Nature* **366**, 461–464 (1993).
- Gems, D. et al. Two pleiotropic classes of *daf-2* mutation affect larval arrest, adult behaviour, reproduction and longevity in *Caenorhabditis elegans*. *Genetics* **150**, 129–155 (1998).
- Conradt, B. & Horvitz, H. R. The *C. elegans* protein EGL-1 is required for programmed cell death and interacts with the Bcl-2-like protein CED-9. *Cell* **93**, 519–529 (1998).
- Duret, L., Gues, N., Peitsch, M. & Bairoch, A. New insulin-like proteins with atypical disulfide bond pattern characterized in *C. elegans* by comparative sequence analysis and homology modeling. *Genome Res.* **8**, 348–353 (1998).
- Peschke, E., Peschke, D., Hammer, T. & Csernus, V. Influence of melatonin and serotonin on glucose-stimulated insulin release from perfused rat pancreatic islets in vitro. *J. Pineal Res.* **23**, 156–163 (1997).
- Breum, L., Bjerre, U., Bak, J. F., Jacobsen, S. & Astrup, A. Long-term effects of fluoxetine on glycemic control in obese patients with non-insulin-dependent diabetes mellitus or glucose intolerance: influence on muscle glycogen synthase and insulin receptor kinase activity. *Metabolism* **44**, 1570–1576 (1995).

- Avery, L. The genetics of feeding in *Caenorhabditis elegans*. *Genetics* **133**, 897–917 (1993).
- Lakowski, B. & Hekimi, S. The genetics of caloric restriction in *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA* **95**, 13091–13096 (1998).
- Nonogaki, K., Strack, A. M., Dallman, M. F. & Tecott, L. H. Leptin-independent hyperphagia and type 2 diabetes in mice with a mutated serotonin 5-HT_{2C} receptor gene. *Nature Med.* **4**, 1152–1156 (1998).
- Barzilai, A., Kennedy, T. E., Sweatt, J. D. & Kandel, E. R. 5-HT modulates protein synthesis and the expression of specific proteins during long-term facilitation in Aplysia sensory neurons. *Neuron* **2**, 1577–1586 (1989).

Acknowledgements

We thank P. Sternberg, S. Nurrish and J. Kaplan for advice on serotonin regulation of behaviour; members of the Ruvkun lab for critical reading of the manuscript. R. Lee for help with graphics; G. Sandoval for some strain characterizations; A. Lander for advice on his microscope and software, and especially R. Lints and S. Emmons for analysing the dopamine accumulation in *tph-1(mg280)*. J.Y.S. was supported by a fellowship from the National Institutes of Health, M.V. by a fellowship from the DAAD German Academic Exchange Service, C.M.L. by NSF. This work was supported by grants from Hoechst AG to G.R., and grants from NIH to Y.S. Some nematode strains were supplied by the *Caenorhabditis* Genetics Center supported by the NIH and the University of Minnesota.

Correspondence and requests for materials should be addressed to G.R. (e-mail: ruvkun@frodo.mgh.harvard.edu).

Evidence for stabilizing selection in a eukaryotic enhancer element

Michael Z. Ludwig*, Casey Bergman*, Nipam H. Patel† & Martin Kreitman*

* Department of Ecology and Evolution, University of Chicago, 1101 E. 57th Street, Chicago, Illinois 60637, USA

† Department of Organismal Biology and Anatomy, and Howard Hughes Medical Institute, MC1028, N-101, 5841 South Maryland Avenue, Chicago, Illinois 60637, USA

Eukaryotic gene expression is mediated by compact *cis*-regulatory modules, or enhancers, which are bound by specific sets of transcription factors¹. The combinatorial interaction of these bound transcription factors determines time- and tissue-specific gene activation or repression. The *even-skipped* stripe 2 element controls the expression of the second transverse stripe of *even-skipped* messenger RNA in *Drosophila melanogaster* embryos, and is one of the best characterized eukaryotic enhancers^{2–4}. Although *even-skipped* stripe 2 expression is strongly conserved in *Drosophila*, the stripe 2 element itself has undergone considerable evolutionary change in its binding-site sequences and the spacing between them. We have investigated this apparent contradiction, and here we show that two chimaeric enhancers, constructed by swapping the 5' and 3' halves of the native stripe 2 elements of two species, no longer drive expression of a reporter gene in the wild-type pattern. Sequence differences between species have functional consequences, therefore, but they are masked by other co-evolved differences. On the basis of these results, we present a model for the evolution of eukaryotic regulatory sequences.

Multiple binding sites for each of four transcription factors, the activators bicoid (*bcd*) and hunchback (*hb*), and the repressors Kruppel (*Kr*) and giant (*gt*), have been physically localized to the stripe 2 element (S2E) (ref. 5). Genetic and experimental evidence indicates that concentration gradients of bicoid and hunchback can activate *eve* in a broad domain, whereas the more localized expression of the S2E is determined by the repressors giant anteriorly and Kruppel posteriorly^{5–7}. Experimental elimination, addition or augmentation of both repressor and activation sites produces predictable changes in reporter-gene expression, and provides evidence for the mechanisms of enhancer function^{5–7}. This model emphasizes the functional importance of binding-site sequences, as well as the

in transformed embryos at a time point and location along the A–P axis that is indistinguishable from native *eve* expression (Fig. 1b, c, e, f). This result indicates either that the mutational differences between these S2Es are functionally unimportant (to the level of resolution allowed by the experimental procedure), or that the changes, if functional, balance one another in such a way as to produce no net functional change in expression.

We predicted that any chimaeric enhancer that contains the conserved sequences of the S2E would have wild-type function if these sequences only are important for *eve* stripe 2 expression. However, if substitutional differences between lineages are functional then it should be possible to create chimaeric sequences that disrupt wild-type function. To test these alternatives, we created two complementary chimaeric enhancers in which the distal or the proximal parts of *D. melanogaster* S2E were substituted with corresponding parts of the *D. pseudoobscura* enhancer (Fig. 1d, g, j, m). The break lies within a conservative block of seven bases midway through the S2E (Fig. 1a, Block 1), which allowed us to create a 'seamless' connection between homologous parts in the chimaeric constructs. The chimaeric enhancers are designated S2E(m1-p2) and S2E(p1-m2), where (x-y) corresponds to species x distal (5') segment connected to species y proximal (3') segment.

In separate experiments involving multiple independent transformants of each chimaeric construct, we found consistent evidence for a posterior shift of about the width of two cells in reporter-gene expression in the S2E(m1-p2) construct (Fig. 1k, l). In some embryos, and in some locations along the circumference of the stripe, this posterior expression appears to involve an expansion of the stripe rather than a simple shift. These two phenotypic defects, posterior shifting and expansion, suggest both a change in the sensitivity to the bcd/hb activation gradients, and a reduction in the overall effectiveness of Kr repression of the posterior stripe margin. The absence of a putative *D. pseudoobscura*-specific Kr site (Fig. 1g, asterisk) in the S2E(m1-p2) chimaera, on the basis of binding-site prediction, provides a testable hypothesis to explain the posterior shift in reporter-gene expression. The complementary construct, S2E(p1-m2), does not show the same defect. Instead, it appears to be defective in that it leads to subtle expansion of both the anterior and posterior borders of the stripe compared with native *eve* expression (Fig. 1h, i).

To explain these results, we propose that stabilizing selection has maintained phenotypic constancy for *eve* expression but has allowed mutational turnover of functionally important sites. Stripe 2 expression can be viewed as a quantitative character in which stabilizing selection has preserved *eve* expression to a specific band of cells and a particular time in embryogenesis. Mutational changes in the element, including some base substitutions within binding sites and short insertions or deletions in the spacer regions between binding sites, are proposed to have only weak functional (and therefore selective) effects. Theoretical models of phenotypic traits under stabilizing selection show that nearly neutral variation (that is, slightly deleterious and advantageous mutations) reaches fixation at an appreciable rate by the process of genetic drift^{13–15}. The model implies that each species lineage will differ by many functionally compensatory mutations. Such a pattern of substitution, we predict, will be a common theme in *cis*-regulatory evolution.

Our results have implications for understanding changes in gene expression and morphological evolution. If our model of enhancer evolution is correct, then weakly selected mutations in regulatory elements will be present in natural populations and will be available for directional change by natural selection. Consistent with this prediction, two traits in *Drosophila* that differ between species but that are likely to be under stabilizing selection within a species—abdominal bristle number and wing morphology—are reported to have significant associations between quantitative trait variation and nucleotide polymorphism in noncoding regions of candidate loci^{16,17}. But as our results indicate, selection can maintain func-

tional conservation of gene expression for long periods of evolutionary time despite binding site turnover, which may make it difficult to identify homologous elements in different species groups by sequence comparison alone. □

Methods

Binding-site prediction

Samples of *Drosophila* DNase 1 footprinted sequences were compiled from the literature for bcd, hb and Kr. Only five sites have been established for gt and thus a position weight matrix (PWM) was not built for this factor. We used the predictive update version of the Gibbs sampler to produce an unbiased local alignment and corresponding binding-site usage matrix¹⁸. We computed the likelihood ratio $\text{Log}(L) = \sum_i \text{Log}[P_{ij}/P_i]$ for all overlapping windows, where P_{ij} is the frequency of base i at position j of the binding-site usage matrix, P_i is the frequency of base i in the background nucleotide composition, and summation is over the width of the binding site. For bcd we used 53 sites from 8 targets; for Kr we used 42 sites from 10 targets; and for hb we used 96 sites from 15 targets (available on request from C.B.). This method successfully identifies all of the experimentally verified binding sites for these three factors (Fig. 1).

Cloning, P-element-mediated transformation and whole-mount *in situ* hybridization

Initial plasmids containing *eve* stripe 2 enhancer regions from *D. melanogaster* and *D. pseudoobscura*, and P-element transformation vector, cloning, amplification, sequencing, P-transformation and whole-mount *in situ* hybridization, analysis of enhancer stripe 2 expression using a reporter gene, and alignment of DNA sequences have been described⁹.

Construction of transgenes

The enhancer regions containing the *eve* stripe 2 element for *D. melanogaster* and *D. pseudoobscura* species, and enhancer halves used for making chimaeric constructs were obtained by PCR from the corresponding plasmids⁹. The fragment containing the S2E(m) element was amplified using primers ME Asp + (5'-aaaaggtactgcataacaatggaaccga-3') and ME Pst - (5'-aaaactgcagtagtcggttaattcggtt-3'). The fragment containing the S2E(p) element was amplified using primers PSE Asp + (5'-aaaaggtactgcataacaatggaaccga-3') and PSE Pst - (5'-aaaactgcagtagtcggttaattcggtt-3'). The fragment S2E(m1) was amplified using primers ME Asp + and ME 1 bl - (5'-agggtccctcgatcgctct-3'). The fragment S2E(m2) was amplified using primers ME 2 bl + (5'-ggactataatgcacacga-3'), and ME Pst-. The fragment S2E(p1) was amplified using primers PSE Asp+ and PSE 1 bl - (5'-agggtccctcgatcgctct-3'). The fragment S2E(p2) was amplified using primers PSE 2 bl + (5'-tgcacacagggtgtctct-3') and PSE Pst-. The chimaeric enhancer S2(m1-p2) was prepared by blunt ligation of S2E(m1) and S2E(p2), and re-amplification with the primers ME Asp+ and PSE Pst-. The chimaeric enhancer S2(p1-m2) was prepared by blunt ligation of S2E(p1) and S2E(m2), and re-amplification with the primers PSE Asp+ and ME Pst-. The primers contained the restriction sites for Asp718 and Pst1, respectively, at their 5' and 3' ends. After digestion with these enzymes, the PCR fragments were cloned into Asp718 and Pst1 sites of the plasmid. Entirely correct inserts were identified for further use by sequencing independent clones. The spacing between the 'experimental' stripe 2 element and the 'control' stripe 3 + 7 element is sufficient to assure independent function of the two enhancers¹². The stripe 3 + 7 *lacZ* expression allows us to control for position effects in expression due to the site of insertion in the genome of different constructs, and it also allows us to precisely control for the time point in development when expression patterns in individual embryos are being compared⁹.

Received 2 July; accepted 23 November 1999.

1. Arnone, M. I. & Davidson, E. H. The hardwiring of development: organization and function of genomic regulatory systems. *Development* **124**, 1851–1864 (1997).
2. Frasch, M. & Levine, M. Complementary patterns of *even-skipped* and *fushi tarazu* expression involve their differential regulation by a common set of segmentation genes in *Drosophila*. *Genes Dev.* **1**, 981–995 (1987).
3. Goto, T. P., Macdonald, P. & Maniatis, T. Early and late periodic patterns of *even-skipped* expression are controlled by distinct regulatory elements that respond to different spatial cues. *Cell* **57**, 413–422 (1989).
4. Harding, K., Hoey, T., Warrior, R. & Levine, M. Autoregulatory and gap response elements of the *even-skipped* promoter of *Drosophila*. *EMBO J.* **8**, 1205–1212 (1989).
5. Stanoevic, D., Small, S. & Levine, M. Regulation of a segmentation stripe by overlapping activators and repressors in the *Drosophila* embryo. *Science* **254**, 1385–1387 (1991).
6. Small, S., Blair, A. & Levine, M. Regulation of *even-skipped* stripe 2 in the *Drosophila* embryo. *EMBO J.* **11**, 4047–4057 (1992).
7. Arnosti, D., Barolo, S., Levine, M. & Small, S. The *eve* stripe 2 enhancer employs multiple modes of transcriptional synergy. *Development* **122**, 205–214 (1996).
8. Hewitt, G. F. et al. Transcriptional repression by the *Drosophila* giant protein: *cis* element positioning provides an alternative means of interpreting an effector gradient. *Development* **126**, 1201–1210 (1999).
9. Ludwig, M. Z., Patel, N. H. & Kreitman, M. Functional analysis of *eve* stripe 2 enhancer evolution in *Drosophila*: rules governing conservation and change. *Development* **125**, 949–958 (1998).
10. Ludwig, M. Z. & Kreitman, M. Evolutionary dynamics of the enhancer region of *even-skipped* in *Drosophila*. *Mol. Biol. Evol.* **12**, 1002–1011 (1995).
11. Beverly, S. M. & Wilson, A. C. Molecular evolution in *Drosophila* and the higher Diptera II. A time scale for fly evolution. *J. Mol. Evol.* **21**, 1–13 (1984).
12. Small, S., Arnosti, D. N. & Levine, M. Spacing ensures autonomous expression of different stripe enhancers in the *even-skipped* promoter. *Development* **119**, 767–772 (1993).

13. Kimura, M. Possibility of extensive neutral evolution under stabilizing selection with special reference to non-random usage of synonymous codons. *Proc. Natl Acad. Sci. USA* **78**, 5773–5777 (1981).
14. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, 1983).
15. Ohta, T. & Tachida, H. Theoretical study of near neutrality. I. Heterozygosity and rate of mutant substitution. *Genetics* **126**, 219–229 (1990).
16. Long, A. D., Lyman, R. E., Langley, C. H. & Mackay, T. F. Two sites in the Delta gene region contribute to naturally occurring variation in bristle number in *Drosophila melanogaster*. *Genetics* **149**, 999–1017 (1998).
17. Gibson, G. & Hogness, D. S. Effect of polymorphism in the *Drosophila* regulatory gene Ultrabithorax on homeotic stability. *Science* **271**, 200–203 (1996).
18. Lawrence, C. E. *et al.* Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**, 208–214 (1993).
19. Berg, O. G. & von Hippel, P. H. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* **193**, 723–750 (1987).

Acknowledgements

We thank S. Small and J. Comeron for helpful discussions. This work was supported by NSF grants to M.K. and M.Z.L. N.H.P. is a Howard Hughes Medical Institute investigator.

Correspondence and requests for materials should be addressed to M.Z.L. (e-mail: mludwig@midway.uchicago.edu).

Structure of human guanylate-binding protein 1 representing a unique class of GTP-binding proteins

Balaji Prakash*, Gerrit J. K. Praefcke*, Louis Renault, Alfred Wittinghofer & Christian Herrmann

Max-Planck-Institut für Molekulare Physiologie, Otto-Hahn-Strasse 11, 44227 Dortmund, Germany

* These authors contributed equally to this work

Interferon- γ is an immunomodulatory substance that induces the expression of many genes to orchestrate a cellular response and establish the antiviral state of the cell. Among the most abundant antiviral proteins induced by interferon- γ are guanylate-binding proteins such as GBP1 and GBP2 (refs 1, 2). These are large GTP-binding proteins of relative molecular mass 67,000 with a high-turnover GTPase activity³ and an antiviral effect⁴. Here we have determined the crystal structure of full-length human GBP1 to 1.8 Å resolution. The amino-terminal 278 residues constitute a modified G domain with a number of insertions compared to the canonical Ras structure, and the carboxy-terminal part is an extended helical domain with unique features. From the structure and biochemical experiments reported here, GBP1 appears to belong to the group of large GTP-binding proteins that includes Mx and dynamin, the common property of which is the ability to undergo oligomerization with a high concentration-dependent GTPase activity⁵.

Guanylate-binding proteins (GBP1 and 2) were originally identified as proteins from an extract of human fibroblasts treated with interferons, γ -interferon being the most effective, that bind to agarose-bound GMP, GDP and GTP^{1,2}. Smaller guanylate-binding proteins of relative molecular mass 47,000 ($M_r = 47K$)⁶ are also induced by γ -interferon, whereas α - and β -interferon induce antiviral GTP-binding Mx proteins⁷. Human (h)GBP1 is expressed to mediate an antiviral effect against vesicular stomatitis virus and encephalomyocarditis virus⁴. The biochemical properties of GBPs are clearly different from those of Ras-like and heterotrimeric GTP-binding proteins. They bind guanine nucleotides with low affinity (micromolar range), are stable in their absence and have a high turnover GTPase^{3,8}. In addition to binding GDP/GTP, they have the unique ability to bind GMP with equal affinity and hydrolyse GTP not only to GDP but also to GMP³. As a first step towards under-

standing the biochemistry and biology of GBPs, we have determined the three-dimensional structure of hGBP1. Furthermore, we show nucleotide-dependent oligomerization of hGBP1 and concentration dependence of its GTPase reaction rate.

We determined the structure of full-length, histidine(His₆)-tagged hGBP1 in the absence of nucleotide to 1.8 Å. The final model comprises residues 6–583 and 341 water molecules (Fig. 1a, b), with some poorly defined, probably mobile loops (dashed lines). The crystallographic data are summarized in Table 1. The structure can be divided into a compact, globular α , β -domain (6–278), which we term the LG (Large G) domain, and an elongated, purely α -helical domain. The domains are connected by a short intermediate region consisting of one α -helix and a short two-stranded β -sheet. The connecting region is not an independent domain; it is packed, via helix α 6, against the β 1/ α 1 region of the LG domain, away from the presumed nucleotide-binding site (see below). It could be involved in stabilizing the relative location of the two domains against each other. The helical domain is composed of seven helices, which extend 90 Å away from the LG domain.

The LG domain of GBPs contains the conserved sequence elements of GTP-binding proteins with modifications. Originally the N/TKxD motif was believed to be absent in the GBPs². An Asp-Asn mutation can produce a change in specificity from guanine to xanthine nucleotides in many GTP-binding proteins such as EF-Tu⁹. As an Asp-Asn mutation in the ¹⁸¹TLRD¹⁸⁴ motif of hGBP1 behaves similarly, it is postulated that Asp 184 should bind the guanine base through a bidentate hydrogen bond⁸. It was thus expected that GBP1 would contain the Ras G domain or a variation thereof. The structure shows that the 278-residue globular domain largely resembles the canonical architecture of Ras (~170 residues), allowing for additions and insertions (labelled I). Superimposing the structures of hGBP1 and Ras-GDP (Fig. 2a) gives a root mean square (r.m.s) deviation of 1.1 Å for 112 common C α atoms. The LG domain consists of an eight-stranded β -sheet with six parallel and two anti-parallel strands surrounded by nine helices, whereas Ras contains six β -strands and five helices (Fig. 1b). Using the secondary structure elements of Ras as the basis for the comparison and retaining the corresponding numbering, the additional elements are β 0, α 0 and β -1 on the N-terminal side of the sheet, and helices α 3' (I3) and α 4' (I4) (Fig. 1b). Apart from a short insertion (I1) in switch I, there are comparatively long loop insertions between the ⁹⁷DxxG¹⁰⁰ motif and α 2 (I2), and between β 6 and α 5 (I5) of the canonical G domain, respectively (Figs 1 and 2). Whereas I1, I2 and I5 could be involved in nucleotide binding, I3 and I4, on the opposite side of the protein, apparently mediate contact with the C-terminal helix.

As GBP is stable in the absence of nucleotide, whereas Ras-like and G α GTP-binding proteins are not, it was of interest to investigate the effect of the absence of nucleotide on the structure. As all P-loop-containing proteins¹⁰ bind the β / γ -phosphate of the nucleotide in a similar manner, and as the role of Asp 184 in binding the guanine base is similar to that of the Asp of the canonical N/TKxD motif, we can locate the nucleotide-binding site of hGBP1 using the RasGDP-hGBP1 overlay (Fig. 2b). From this comparison we can also see that, although part of the binding site is more accessible to the solvent than in Ras-nucleotide complexes, part of the polypeptide chain is in a position that interferes with nucleotide binding. Perhaps owing to the absence of nucleotide, the polypeptide chain around the binding site is mobile, as no electron density is visible for residues 69–72 (I1) in the region analogous to switch I, residues 190–193 close to the ¹⁸¹TLRD¹⁸⁴ motif and residues 244–257 in I5, close to the SAK/L motif, which is conserved only in the Ras family and is absent in GBPs.

The (phosphate-binding) P loop¹⁰, residues 45–52, adopts a structure different from that of the Ras-nucleotide complexes. The invariant lysine residue of the P loop does not interact with the main-chain carbonyls for stabilization. Instead, in hGBP1 the

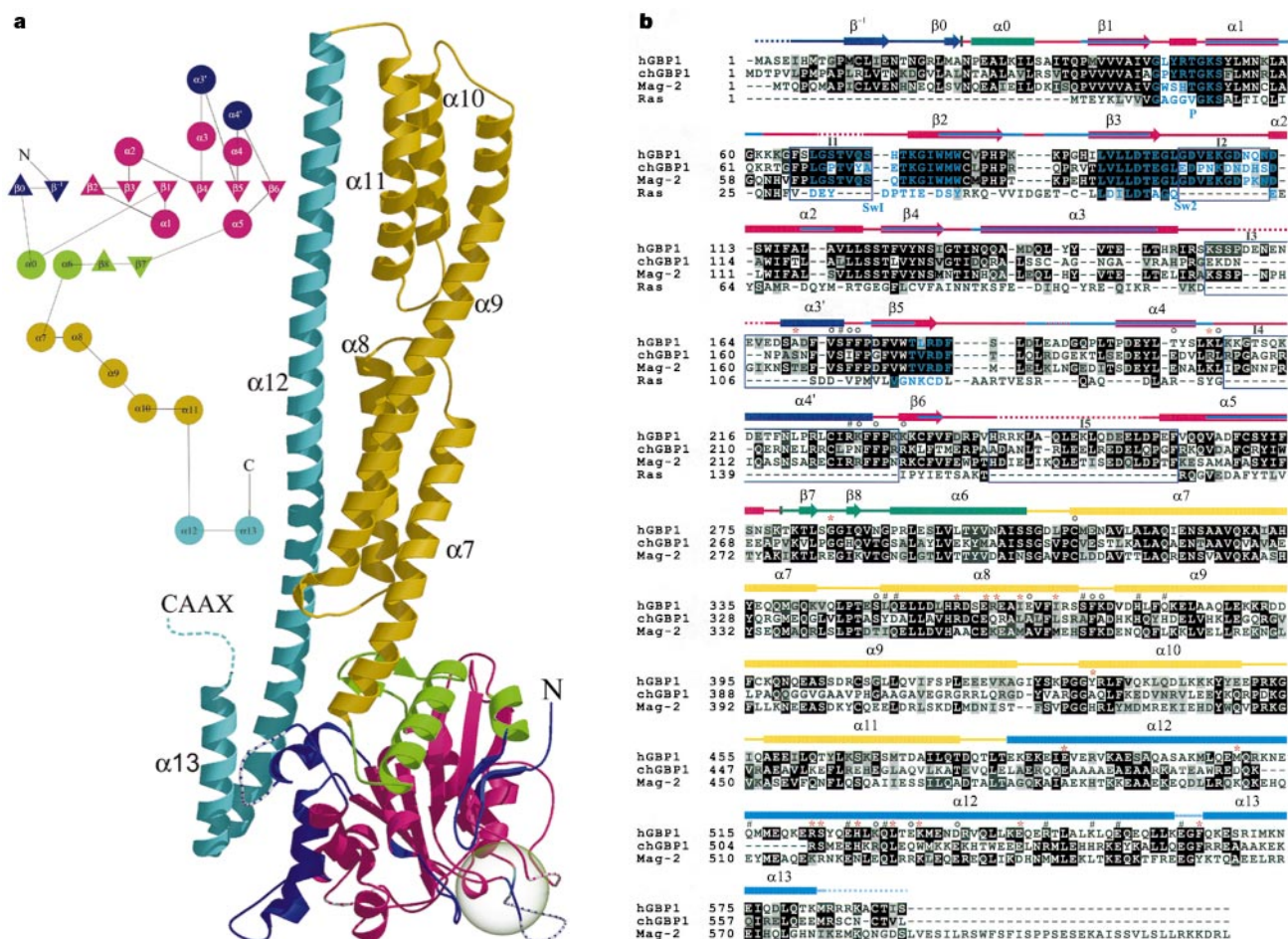


Figure 1 Primary, secondary and tertiary structure of hGBP1. **a**, A model of the tertiary structure of hGBP1 presented as a ribbon, where the LG domain is in purple, the connecting region in green, the helical domain in yellow and $\alpha 12/\alpha 13$ in cyan. Insertions, marked I2–I5 in **b**, are in violet. Dashed lines indicate disordered regions in the molecule. The tentative nucleotide-binding area, identified by a Ras–GBP overlay, is indicated by a sphere with radius 7 Å. The topology is shown schematically using the same colour code. **b**, Sequence alignment of hGBP1 (Swissprot accession no. P32455) with Mag-2 (EMBL

loop is stabilized by interactions with the region analogous to switch II, involving hydrogen bonds between Tyr 47 (backbone N) and Asp 103, Lys 51 and Thr 98. Furthermore, the structure is not suited for nucleotide binding as the phosphates would clash with Tyr 47. The region corresponding to switch I in Ras is disordered in hGBP1. Thr 75 appears to be analogous to Thr 35 in Ras, but is 5 Å away in the overlay (Fig. 2b). The $^{97}\text{DxxG}^{100}$ motif of hGBP1 superimposes well with that of switch II in Ras. D184 is 6 Å away from the corresponding D119 of the canonical N/TKxD motif and would have to move accordingly to occupy a similar position in the nucleotide-bound form. In general, it appears that the guanine nucleotide-binding site is partly open (Fig. 2a, b) such that the incoming nucleotide would enter the binding site base first and would then, after a corresponding conformational change, bind into the phosphate-binding area, as suggested for Ras by the structure of the Ras–Sos complex¹¹.

Ras proteins have an intrinsic GTPase reaction rate in the order of $0.001\text{--}0.1\text{ min}^{-1}$, whereas hGBP1 (ref. 8) has a rate of up to 80 min^{-1} (see below). The catalytic mechanism for Ras and G α proteins involves a glutamine (Gln 61 in Ras) that stabilizes the transition state and is itself stabilized by GAP in the GAP-catalysed mechanism¹². As Gly 60^{Ras} and Gly 100^{hGBP1} are very close to each other (Fig. 2b), we would predict that Gln 61 of Ras is structurally homologous to Leu 101 in hGBP1. Considering the importance of

accession no. M81128) from mouse and chicken GBP1 (EMBL accession no. X92112), with the secondary structure assignment as determined using the programme DSSP²⁹, with the same colour code as in **a**, and aligned with the secondary structure of Ras-GppNHp³⁰ (light blue lines). Contacts between helix $\alpha 12$ and the rest of the protein are indicated: asterisk, direct; circle, water-mediated polar interactions; hash sign for both. For brevity, sequences with lowest homology have been chosen.

Gln 61 in the intrinsic and GAP-catalysed GTPase reaction of Ras, we conclude that the chemical mechanism of hydrolysis of GTP by hGBP1 is different from that of Ras. A similar hydrophobic residue corresponding to Gln 61 of Ras has also been identified in dynamin and Mx proteins. Discounting the influence of oligomerization on the rate of the GTPase reaction, only insertion I2 following Leu 101 appears to be close enough to the phosphate-binding site to be involved in catalysis. This $^{103}\text{DxEKGD}^{108}$ motif is conserved in GBPs and contains charged residues that could have a catalytic role, either similar to that of the Arg in G α proteins or GAPs¹² or as a catalytic base.

The helical domain of hGBP1 can be considered to consist of two three-helix bundles, formed by helices $\alpha 7$, $\alpha 8$, $\alpha 9$ (residues 311–403) and helices $\alpha 9$, $\alpha 10$, $\alpha 11$ (404–482), where helix $\alpha 9$ is common to both (Fig. 1a). Adjacent to and covering these two bundles is a very long helix $\alpha 12$ that reaches back to the LG domain. This helix, which is predicted from the sequence to be a coiled-coil structure, is 78 residues long and stretches over 118 Å. At its C-terminal end there is a helical turn leading into another short helix, $\alpha 13$, which makes a coiled-coil type of interaction with $\alpha 12$. The last eight residues, including the prenylation-recognition Caax motif, not modified owing to recombinant expression in *E. coli*, are not visible in the structure.

The two three-helix bundles give the molecule an elongated shape. The core of the helical domains is formed by hydrophobic

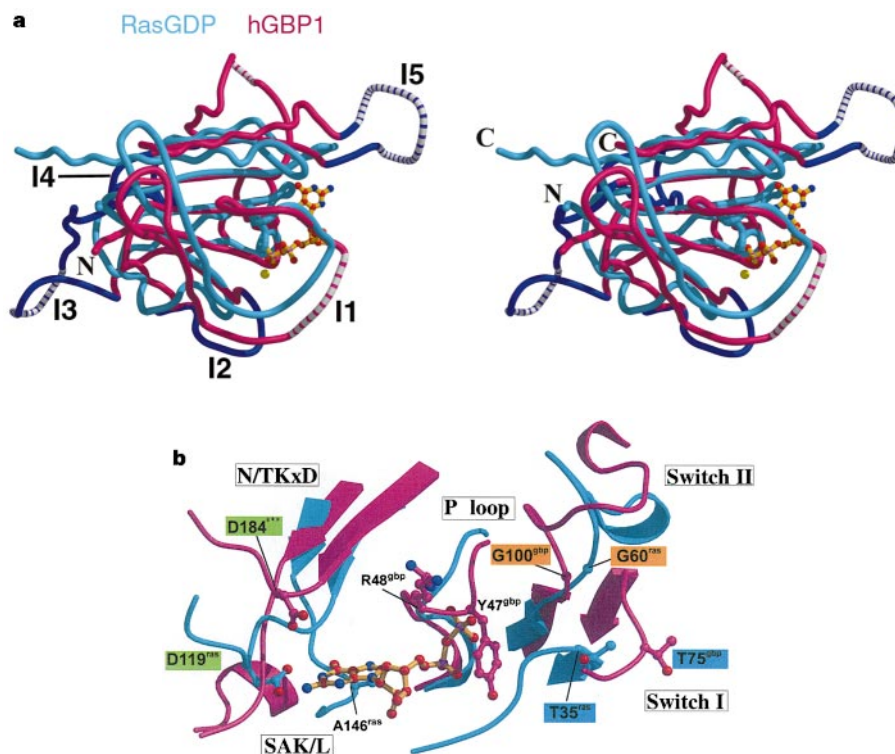


Figure 2 Comparison of hGBP1 and Ras structures. **a**, Superimposition of the LG domain of hGBP1 with the G domain of Ras in complex with GDP (PDB accession no. 1Q21) as a stereo view. N-terminal residues 1–36 of hGBP1 up to $\beta 1$ have been omitted for clarity. The colour code is as in Fig. 1; Ras is in cyan. **b**, Putative location of nucleotide-binding site in hGBP1. The regions of hGBP1 potentially involved in binding the guanine nucleotide

are shown as obtained from a structural superimposition of RasGDP (in cyan) with the corresponding regions in hGBP1 (purple), highlighting functionally important residues necessary for binding and conformational change as balls or in ball-and-stick. Whereas Gly 60^{Ras} overlays very well with Gly 100^{hGBP1}, residues D119/D184 and T35/T75 do not.

Table 1 Crystallographic data statistics

Data collection and phase determination by MIRAS

Crystal space group: C2 ($a = 155.47 \text{ \AA}$, $b = 41.03 \text{ \AA}$, $c = 121.00 \text{ \AA}$, $\beta = 109.42^\circ$)

Parameter	Native*	Derivatives				
		KAu(CN) ₂	EMP-1	EMP-2	K ₂ Pt(CN) ₆	K ₂ WO ₄
Soaking concentration and time		0.5 mM 48 h	0.1 mM 6.5 h	†	2.5 mM 1.5 h	0.2 mM 1 h
Resolution (Å)	39.5–1.8	17–2.4	28–2.47	39–2.5	31–2.47	29–2.44
High-resolution shell	1.9–1.80	2.44–2.40	2.5–2.47	2.6–2.5	2.55–2.47	2.47–2.44
X-ray source	DESY(BW6)	DESY(BW6)	RA‡	RA‡	DESY(BW6)	RA‡
Wavelength (Å)	1.07, 1.54§	1.035	1.54	1.54	1.067	1.54
Completeness (%)	99.0 (95.1)	92.9 (59.6)	92.8 (64.8)	87.5 (68.4)	95.3 (90.7)	95.4 (69.5)
Unique reflections	66,520	26,371	24,570	23,419	26,220	27,232
Redundancy	5.7 (3.0)	2.0	3.4	2.9	4.5	4.3
$\ R_{\text{sym}}\ $ (%)	5.9 (29.9)	5.6 (9.5)	7.2 (23.4)	7.9 (25.2)	6.8 (17.1)	10.3 (34.0)
I/σ	17.1 (3.7)	13.1 (6.0)	14.6 (2.5)	12.2 (2.7)	16.3 (6.6)	13 (2.7)
$\ R_{\text{iso}}\ $ (%)		26.1	30.9	23.2	19.1	19.4
Number of sites		4	7	5	5	4
# R_{cullis} (ac/c)		0.81/0.86	0.62/0.76	0.63/0.75	0.85/0.91	0.94/0.94
Anom. R_{cullis}		0.94	0.94	0.96	0.97	–
* R_{work} (ac/c)		1.2/0.81	2.2/1.3	2.17/1.34	1.05/0.71	0.76/0.54
Overall figure of merit	0.6807					
Refinement statistics						
Resolution	30–1.8 Å			** R_{work}		24.3%
Reflections (work set/test set)	59798/6722			** R_{free}		27.6%
Protein atoms	4362			r.m.s.d. bond length		0.006 Å
No. of water molecules	341			r.m.s.d. bond angles		1.1°
Average B (Å ²)	35.4					

Values in parentheses correspond to the highest resolution shell; ac/c refer to acentric and centric data. EMP, ethylmercurypyrophosphate.

† The derivative EMP-2 was obtained by pre-soaking the protein in EMP for 3 h before setting up crystallization.

* Cryoprotectant used: 7.5% PEG6000, 20% glycerol, 15% xylitol (w/v) in 100 mM MES buffer, pH 6.0.

‡ RA: Rotating anode.

§ Native data set was obtained by merging the data collected at DESY(BW6) with the data collected on a rotating anode to a resolution of 2.4 Å with a redundancy of 5.9.

$\|R_{\text{sym}}\| = \sum_i \sum_j |I(h) - I(h)| / \sum_i \sum_j I(h)$, where $I(h)$ and $\|I(h)\|$ are the i th and mean measurements of the intensity of reflection h .

$\|R_{\text{iso}}\| = \sum_i |F_P - F_{PH}| / \sum_i |F_P|$, where F_{PH} and F_P are the observed structure factor amplitudes for the derivative and native data sets, respectively, of the reflection h .

$R_{\text{cullis}} = \sum_i |F_{PH} \pm F_P| - F_{PH} / \sum_i |F_{PH} \pm F_P|$, where F_{PH} is the calculated structure factor for heavy atoms.

* $R_{\text{work}} = (\sum_i |F_o - F_c| / \sum_i |F_o|)^{1/2}$, where $F_{o, \text{obs}}$ and $F_{o, \text{calc}}$ are the observed and calculated structure factor amplitudes for the derivative.

** $R_{\text{free}} = \sum_i |F_o - F_c| / \sum_i |F_o|$, where F_o and F_c are the observed and calculated structure factor amplitudes of reflection h , respectively.

R_{free} is the same as R_{work} , but calculated on 10% of data set aside for wARP and for refinement.

r.m.s. deviations (r.m.s.d.) are given as deviations from ideal values.

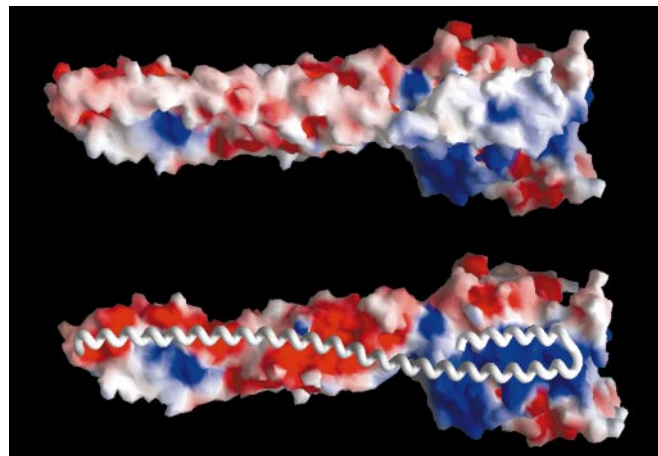


Figure 3 Interaction of the C-terminal helix motif $\alpha 12/13$ with the helical and the LG domains. The electrostatic surface potential shows that the highly charged regions of the helical and LG domains are masked by an $\alpha 12/13$ motif, as indicated in the lower panel by showing $\alpha 12/13$ in worm representation.

residues, whereas the charged amino acids are exposed towards but interact only weakly with the residues from the long helix $\alpha 12$. Thus, $\alpha 12$ has almost no contact with the second (more remote) helical bundle (a single hydrogen bond between Glu 490 and Tyr 433, Fig. 1b) whereas it forms seven direct side-chain and nine water-mediated contacts with the first. It is stabilized further by four direct and eight water-mediated contacts with insertions I3 and I4 (Fig. 1) of the LG domain. The electrostatic surface potential (Fig. 3) indicates that $\alpha 12$ may mask some of the charged residues exposed by the two helical bundles. In conclusion, the helical domain appears to actually consist of two subdomains.

Mx and dynamin are commonly grouped into a separate class of large GTP-binding proteins⁵. Proteins of this family have several variations in their domain structure, but they all possess at least a 'GTPase' domain of ~300 residues, a 'middle' or 'assembly' domain (150–200 residues) and a 'GED' domain (100 residues). These values are close to the sizes of the LG, helical bundle and $\alpha 12/13$ domains of hGBP1, respectively. Biochemically, they share common properties such as relatively low affinity for nucleotides (K_m 10–500 μ M), the ability to form oligomers and a high intrinsic rate of GTP hydrolysis (k_{cat} 2–100 min^{-1})^{5,7}. It was thus interesting to find out whether hGBP1 has to undergo multimerization to hydrolyse GTP rapidly, as observed for dynamin. Figure 4a shows that the specific GTPase activity does increase at least eight-fold with increasing concentrations under the conditions employed. This clearly indicates a cooperative mechanism of GTP hydrolysis by hGBP1, similar to that observed for dynamin¹³. Note, however, that for technical reasons the data do not allow us to estimate the GTPase reaction rate at lower concentrations.

The homology to the dynamin family is further corroborated by demonstrating nucleotide-dependent oligomerization (Fig. 4b). The elution profiles from standardized analytical size-exclusion chromatography show that hGBP1 is monomeric in the absence and in the presence of GDP, whereas it shows higher molecular mass when bound to the non-hydrolysable GTP analogue GppNHp. The molecular mass is higher again for hGBP1 in complex with GDP and aluminium fluoride which together are believed to mimic the transition state of GTP hydrolysis^{8,12}. It seems that, for efficient GTP hydrolysis to occur, higher order structures have to be formed, as suggested for dynamin^{13–15}. The dependence of nucleotide hydrolysis on oligomerization has been proposed to be important for dynamin's ability to assemble around the necks of clathrin-coated vesicles and regulate receptor-mediated endocytosis. A role for the oligomerization-dependent hydrolysis of GTP by hGBP1 in interferon-mediated cellular functions remains to be established.

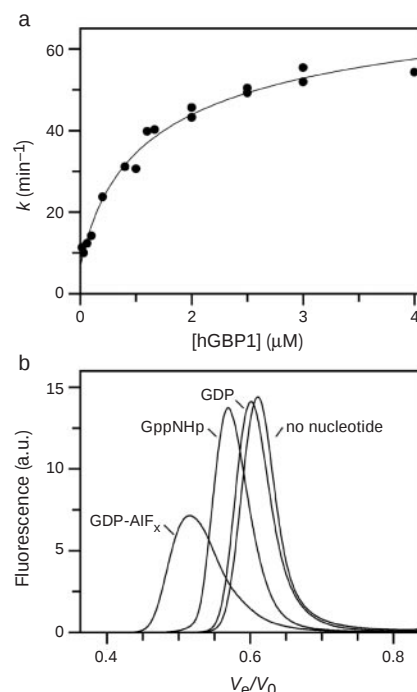


Figure 4 Cooperative GTP hydrolysis and nucleotide-dependent multimerization of hGBP1. **a**, Concentration dependence of the specific GTPase activity of hGBP1. The data were fitted to a model involving dimer formation yielding 0.6 μ M for the apparent dimerization constant of hGBP1. **b**, Size-exclusion chromatography of GDP- AlF_x , GppNHp-, GDP-bound and nucleotide-free hGBP1.

In addition to the biochemical experiments, structural considerations lead us to propose that hGBP1 belongs to the dynamin family of large GTP-binding proteins. Secondary structure predictions suggest a helical and/or coiled-coil structure for the middle and GTPase effector (GED) domains of Mx and dynamin, just as the secondary structure prediction for $\alpha 7$ – $\alpha 12$ of hGBP1 is in tune with the current structure. Limited proteolysis experiments¹⁶ and yeast two-hybrid studies¹⁷ have shown that the C-terminal end of Mx, which is required for fast GTP hydrolysis, contacts the nucleotide-binding domain and the middle domain just as helix $\alpha 12$ folds back onto the LG domain and the helical bundles in the structure of hGBP1. A similar pattern of interactions has also been observed for dynamin where the binding of the GED domain to both the middle and the LG domain has been shown using various techniques^{18,19}. Similarly, the internal GAP-related GED domain of dynamin is implicated not only in GTP hydrolysis but also in the regulation of assembly: two processes that are intimately coupled^{14,20}. With the structure of hGBP1 now available, we can design experiments to test the hypothesis of a common structural make-up for these proteins. We would also investigate the mechanism of GTP hydrolysis in GBPs, the unique nature of the resulting products (GMP and GDP) and whether an internal helical GED domain with properties similar to that of dynamin exists in GBPs. Furthermore, the role of the C-terminal end in the biological function of GBPs in the antiviral defence should be investigated, as the functionally related Mx protein requires its C-terminal end to interact with thogoto virus nucleocapsids²¹. □

Methods

Protein preparation

hGBP1 with an N-terminal His₆ tag was expressed from a pQE9 vector in *E. coli* strain BL21(DE3) as described⁸. The GTPase activity of hGBP1 was measured at 37 °C in a buffer containing 50 mM Tris/HCl, pH 8.0, 5 mM MgCl₂ and 300 μ M GTP. Aliquots were taken at adequate time intervals and the progress of GTP hydrolysis was analysed by reversed phase chromatography (C-18, Hypersil)⁸. Rates were derived from a linear fit to the initial reaction (<30% GTP hydrolysed).

We analysed the states of multimerization of hGBP1 by size-exclusion chromatography (GFC-1300, Supelco). hGBP1 at 20 μ M was preincubated for 15 min with 200 μ M nucleotide and the column was equilibrated with 50 mM Tris/HCl pH 8.0, 5 mM MgCl₂ and 200 μ M of the corresponding nucleotide, or in addition with 300 μ M AlCl₃ and 10 mM NaF. Elution volumes (V_e/V_0) were compared to those of standard proteins.

Crystallography

We used full-length his₆-tagged hGBP1 (40 mg ml⁻¹) to grow crystals in hanging drops over a reservoir solution of 5% PEG6000 (polyethylene glycol), 100 mM MES-NaOH, pH 6.0 at 20 °C. Owing to a cloning artefact, we used the Q507H mutation of hGBP1, which does not alter the biochemical properties. We collected a high-resolution native data set to 1.8 Å resolution at 100K at BW6 beamline (DESY) and processed with XDS: better completeness and redundancy were achieved by merging this data with a 2.4 Å native data set collected on a rotating anode. The statistics of the native and derivative data used for phasing are shown in Table 1. All crystallographic calculations were performed using the CCP4 suite of programs²⁰ unless stated otherwise. Heavy atom sites, initially identified by SOLVE²² and later by difference-Fourier maps, were refined and phase information was generated using MLPHARE. We improved the MIRAS phases by solvent flattening and histogram matching with DM. However, three of the derivatives shared common major sites, as reflected in the high figure of merit shown in Table 1. The map thus obtained posed difficulties in tracing the polypeptide chain. We therefore used wARP²³ to improve the MIRAS phases and extend the map to 1.8 Å resolution. The resulting maps considerably facilitated map interpretation and model building. At this stage, electron-density maps from two multi-wavelength anomalous dispersion (MAD) experiments with Au and Hg derivatives, performed at beamline BM14 (ESRF), aided map interpretation. The poor quality of crystals used for MAD experiments and non-isomorphism between crystals resulted in poorer maps upon combining the MAD and MIRAS phases. We subjected an initial model consisting of 473 residues to iterative processes of building, refinement and phase combination before arriving at a final model containing 548 residues and 341 water molecules. The atomic model was built using program O²⁴, and all refinement, consisting of bulk-solvent correction, positional, torsion angle simulated annealing and B-factor refinement, was carried out with CNS²⁵. Composite simulated annealing omit maps were used to correct or build ambiguous regions of the model. The Ramachandran plot (not shown) depicts 96% of the main-chain torsion angles in the most favoured regions, with no amino acids in the disallowed regions. The current model extends from His 6 to Met 583. The loops comprising residues 63–72, 157–166, 190–193 and 244–256 cannot be seen in the electron density and are presumably disordered. Figures were generated using Molscript²⁶ and Raster3D²⁷. The electrostatic surface potential was generated using GRASP²⁸.

Received 19 October; accepted 7 December 1999.

- Cheng, Y.-S. E., Colonna, R. J. & Yin, F. H. Interferon induction of fibroblast proteins with guanylate binding activity *J. Biol. Chem.* **258**, 7746–7750 (1983).
- Cheng, Y.-S. E., Patterson, C. E. & Staeheli, P. Interferon-induced guanylate-binding proteins lack an N(T)KXD consensus motif and bind GMP in addition to GDP and GTP *Mol. Cell. Biol.* **11**, 4717–4725 (1991).
- Schwemmler, M. & Staeheli, P. The interferon-induced 67-kDa guanylate-binding protein (hGBP1) is a GTPase that converts GTP to GMP *J. Biol. Chem.* **269**, 11299–11305 (1994).
- Anderson, S. L., Carton, J. M., Lou, J., Xing, L. & Rubin, B. Y. Interferon-induced guanylate-binding protein-1 (GBP-1) mediates an antiviral effect against vesicular stomatitis virus and encephalomyocarditis virus *Virology* **256**, 8–14 (1999).
- van der Blik, A. M. Functional diversity in the dynamin family *Trends Cell Biol.* **9**, 96–102 (1999).
- Boehm, U. *et al.* Two families of GTPases dominate the complex cellular response to IFN- γ *J. Immunol.* **161**, 6715–6723 (1998).
- Staeheli, P., Pitossi, F. & Pavlovic, J. Mx proteins: GTPases with antiviral activity *Trends Cell Biol.* **3**, 268–272 (1993).
- Praefcke, G. J. K., Geyer, M., Schwemmler, M., Kalbitzer, H. R. & Herrmann, C. Nucleotide-binding characteristics of human guanylate-binding protein 1 and identification of the third canonical GTP-binding motif *J. Mol. Biol.* **292**, 321–332 (1999).
- Weijland, A. & Parmeggiani, A. Towards a model for the interaction between elongation factor Tu and the ribosome *Science* **259**, 1311–1314 (1993).
- Saraste, M., Sibbald, P. R. & Wittinghofer, A. The P-loop—a common motif in ATP- and GTP-binding proteins *Trends Biochem. Sci.* **15**, 430–434 (1990).
- Boriack-Sjodin, P. A., Margarit, S. M., Barsagi, D. & Kuriyan, J. The structural basis of the activation of Ras by Sos *Nature* **394**, 337–343 (1998).
- Scheffzek, K., Ahmadian, M. R. & Wittinghofer, A. GTPase activating proteins: helping hands to complement an active site *Trends Biochem. Sci.* **23**, 257–262 (1998).
- Warnock, D. E., Hinshaw, J. E. & Schmid, S. L. Dynamin self-assembly stimulates its GTPase activity *J. Biol. Chem.* **271**, 22310–22314 (1996).
- Sever, S., Muhlberg, A. B. & Schmid, S. L. Impairment of dynamin's GAP domain stimulates receptor-mediated endocytosis *Nature* **398**, 481–486 (1999).
- Stowell, M. H. B., Marks, B., Wigge, P. & McMahon, H. T. Nucleotide-dependent conformational changes in dynamin: evidence for a mechanochemical molecular spring *Nature Cell Biol.* **1**, 27–32 (1999).
- Schwemmler, M., Richter, M. F., Herrmann, C., Nassar, N. & Staeheli, P. Unexpected structural requirements for GTPase activity of the interferon-induced MxA protein *J. Biol. Chem.* **270**, 13518–13523 (1995).
- Schumacher, B. & Staeheli, P. Domains mediating intramolecular folding and oligomerization of MxA GTPase *J. Biol. Chem.* **273**, 28365–28370 (1998).
- Smirnova, E., Shurland, D.-L., Newman-Smith, E. D., Pishvae, B. & van der Blik, A. M. A model for dynamin self-assembly based on binding between three different protein domains *J. Biol. Chem.* **274**, 14942–14947 (1999).

- Okamoto, P. M., Tripet, B., Litowski, J., Hodges, R. S. & Vallee, R. B. Multiple distinct coiled-coils are involved in dynamin self-assembly *J. Biol. Chem.* **274**, 10277–10286 (1999).
- Collaborative Computational project, N.A. The CCP4 suite: programs for protein crystallography *Acta Crystallogr. D* **50**, 760–763 (1994).
- Kochs, G. & Haller, O. GTP-bound Human MxA protein interacts with the nucleocapsids of Thogoto virus (Orthomyxoviridae) *J. Biol. Chem.* **274**, 4370–4376 (1999).
- Terwilliger, T. C. & Berendzen, J. Correlated phasing of multiple isomorphous replacement data *Acta Crystallogr. D* **52**, 749–757 (1996).
- Perrakis, A., Sixma, T. K., Wilson, K. S. & Lamzin, V. S. wARP:improvement and extension of crystallographic phases by weighted averaging of multiple refined dummy atomic models *Acta Crystallogr. D* **53**, 448–455 (1997).
- Jones, T. A. & Kjeldgaard, M. Electron-density map interpretation *Methods Enzymol.* **277**, 173–208 (1997).
- Brunger, A. T. *et al.* Crystallography and NMR system: a new software system for macromolecular structure determination *Acta Crystallogr. D* **54**, 905–921 (1998).
- Kraulis, P. J. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures *J. Appl. Crystallogr.* **24**, 946–950 (1991).
- Merrit, E. A. & Murphy, M. E. P. Raster3D version 2.0. A program for photorealistic molecular graphics *Acta Crystallogr. D* **50**, 869–873 (1994).
- Nicholls, A., Sharp, K. A. & Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons *Proteins* **11**, 281–296 (1991).
- Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features *Biopolymers* **22**, 2577–2637 (1983).
- Pai, E. F. *et al.* Refined crystal structure of the triphosphate conformation of H-ras p21 at 1.35 Å resolution: implications for the mechanism of GTP hydrolysis *EMBO J.* **9**, 2351–2359 (1990).

Acknowledgements

The work was supported by the Deutsche Forschungsgemeinschaft (C.H.) and by Boehringer Ingelheim Fonds (G.J.K.P.). We thank the staff at beamlines BW6, DESY, Hamburg and at BM-14, ESRF, Grenoble for help with data collection. We also thank I. Schlichting, I. Vetter and R. Hillig for discussions and M. Hess for help with figures. A. Beste for help with HPLC and R. Schebaum for secretarial assistance.

Correspondence and requests for materials should be addressed to A.W. (e-mail: alfred.wittinghofer@mpi-dortmund.mpg.de) or to C.H. (e-mail: christian.herrmann@mpi-dortmund.mpg.de). Coordinates have been deposited with the Protein Data Bank (accession no. 1DG3).

Alternative modular polyketide synthase expression controls macrolactone structure

Yongquan Xue & David H. Sherman

Department of Microbiology and Biological Process Technology Institute, University of Minnesota, Box 196, 1460 Mayo Memorial Building, Minneapolis, Minnesota 55455, USA

Modular polyketide synthases are giant multifunctional enzymes that catalyse the condensation of small carboxylic acids such as acetate and propionate into structurally diverse polyketides that possess a spectrum of biological activities^{1,2}. In a modular polyketide synthase, an enzymatic domain catalyses a specific reaction, and three to six enzymatic domains involved in a condensation-processing cycle are organized into a module³. A fundamental aspect of a modular polyketide synthase is that its module arrangement linearly specifies the structure of its polyketide product³. Here we report a natural example in which alternative expression of the pikromycin polyketide synthase results in the generation of two macrolactone structures. Expression of the full-length modular polyketide synthase PikAIV in *Streptomyces venezuelae* generates the 14-membered ring macrolactone narbonolide, whereas expression of the amino-terminal truncated form of PikAIV leads to 'skipping' of the final condensation cycle in polyketide biosynthesis to generate the 12-membered ring macrolactone 10-deoxymethynolide. Our findings provide insight into the structure and function of modular polyketide synthases, as well as a new set of tools to generate structural diversity in polyketide natural products.

Ever since the discovery of a modular polyketide synthase (PKS)^{3,4}, the co-production of and biosynthetic basis for methymycin and pikromycin in *S. venezuelae* ATCC 15439 have remained an enigma. Methymycin and its co-metabolite neomethymycin (Fig. 1) are derived from the 12-membered ring macrolactone 10-deoxymethynolide and are produced in SCM medium⁵, whereas pikromycin and narbonolide (Fig. 1) are derived from the 14-membered ring macrolactone narbonolide and are produced in PGM medium⁶. The apparent structural relationship between 10-deoxymethynolide and narbonolide has drawn research interest on their biosynthesis for the past decade^{7,8}. We have shown that the production of both 10-deoxymethynolide and narbonolide is mediated by a single PKS cluster (*pikA*) in *S. venezuelae*⁹. The *pikA*-encoded PKS is composed of PikAI, PikAII, PikAIII and PikAIV (Fig. 1) modular PKSs, which are similar to EryAI–EryAIII in the erythromycin PKS cluster except that PikAIII and PikAIV each contain a single module, in contrast to the bimodular EryAIII (ref. 3). Moreover, PikAV is an independent thioesterase (TEII) that is distinct from the thioesterase (TE) domain that is located at the carboxy terminus of PikAIV. The modular organization of PikAI–PikAIV predicts the structure of heptaketide 1 (Fig. 1) which can be released and cyclized into narbonolide by the TE domain in PikAIV. The organization of PikAI–PikAIII correlates with the structure of hexaketide 2 (Fig. 1), but termination and cyclization at this stage remains an intriguing issue, especially after genetic evidence (see below) excluded PikAV (TEII) as the determining factor in alternative termination.

The *pikAIV* gene encodes the sixth module in the Pik PKS and is responsible for the final condensation to create heptaketide 1 and narbonolide (Fig. 1); however, the sixth condensation is not necessary for the production of hexaketide 2 (Fig. 1). Therefore, we anticipated different roles for PikAIV and developed two strategies to study its potential function in the production of

narbonolide and 10-deoxymethynolide. The first strategy relied on gene disruption of *pikAIV*, whereas the second strategy involved direct probing of PikAIV from protein extracts of *S. venezuelae* obtained using the alternative culture conditions required to generate the two antibiotics.

Two mutant strains of *S. venezuelae* were created in which PikAIV was disrupted by deleting the C-terminal TE domain and simultaneously introducing a hexahistidine (His₆) sequence into its position for detection of PikAIV. In mutant AX910, an in-frame deletion was engineered to remove the TE domain from PikAIV on the *S. venezuelae* chromosome. In a second mutant (AX912), the TE domain and the downstream TEII gene (*pikAV*) were removed from the bacterial chromosome. As expected, *S. venezuelae* AX912 was devoid of antibiotic production as the mutant lacks thioesterase activities that are necessary to release the polyketide chain from the Pik PKS. AX910 was also devoid of methymycin and neomethymycin production, even though the TEII (*pikAV*) gene was preserved. This result clearly indicated that Pik TEII is not the termination factor in 10-deoxymethynolide formation. It also provided the first evidence that PikAIV (or at least the TE domain within PikAIV) is involved directly in the production of both the 12- and the 14-membered ring macrolactones.

Western blot analysis of PikAIV synthesis corroborated the gene-disruption results. The His₆-tagged PikAIV signal was present in protein extracts obtained from *S. venezuelae* AX910 and AX912 grown under conditions for 10-deoxymethynolide and narbonolide production. The western blot revealed that the PikAIV polypeptide synthesized under culture conditions for 10-deoxymethynolide biosynthesis is different from that generated during culture conditions for narbonolide production. Specifically, the blot showed a single protein band of relative molecular mass 110,000 (*M*_r 110K) from extracts of *S. venezuelae* AX910 and AX912 grown under

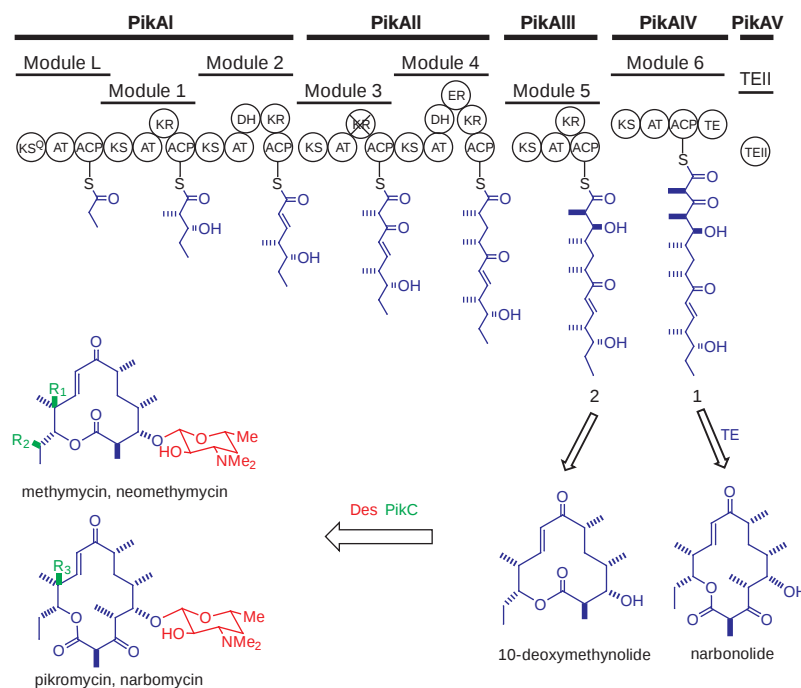


Figure 1 Structure and biosynthesis of methymycin, pikromycin and related compounds in *Streptomyces venezuelae* ATCC 15439. Methymycin: R₁, OH; R₂, H. Neomethymycin: R₁, H; R₂, OH. Pikromycin: R₃, OH. Narbonolide: R₃, H. Polyketide synthase components PikAI, PikAII, PikAIII, PikAIV and PikAV are represented by strings of circles. Each circle represents an enzymatic domain in the Pik PKS system. ACP, acyl carrier protein; AT, acyl-transferase catalysing the transfer of the acyl group from acyl-CoA to form acyl-ACP; KS, ketosynthase catalysing the decarboxylative condensation between two adjacent β -ketoacyl ACPs, which results in an elongated β -ketoacyl chain on the downstream ACP;

KR, ketoreductase reducing β -ketoacyl-ACP into β -hydroxyacyl-ACP; DH, dehydratase converting β -hydroxyacyl-ACP into enoylacyl-ACP; ER, enoyl reductase saturating the α,β double bond; KS⁰, a KS-like domain which is a decarboxylase and serves as a chain initiation factor²⁰ in polyketide biosynthesis; KR with cross, nonfunctional KR; TE, thioesterase releasing the acyl chain from adjacent ACP and presumably cyclizes it into a macrolactone; TEII, type II thioesterase (whose function is obscure). Des represents all eight enzymes for desosamine biosynthesis and transfer⁹, and PikC is the cytochrome P450 monooxygenase responsible for hydroxylation at the R₁, R₂ and R₃ positions²¹.

narbonolide production conditions (Fig. 2a), which is consistent with the predicted TE domain-deleted (His₆-tag replaced) form of *PikAIV*. In contrast, a protein band of ~85K was detected in the extracts obtained from cultures grown using conditions for 10-deoxymethynolide production (Fig. 2b). The 85K band and the 110K band were both occasionally detected in the mutant protein extracts (Fig. 2c), which is consistent with the observation that wild-type *S. venezuelae* occasionally co-produces 12- and 14-membered ring macrolides. Therefore, we concluded that the 85K protein is an N-terminal truncated form of *PikAIV* and is associated with 10-deoxymethynolide (12-membered ring macrolactone) production.

We investigated the generation of the N-terminal truncated form of *PikAIV* (85K band) by mutational analysis. Nucleotide sequence analysis of the *pikAIV* region revealed an alternative translational start site 600 nucleotides downstream of the initial start site. This site has two adjacent GTGs as potential start codons and has a putative ribosome-binding site seven nucleotides upstream from the first GTG (Fig. 2d). Translation from this site predicts an 88K form of *PikAIV*, which is very similar to the observed 85K band in the western blot. When these two GTG codons were mutated to GTC, 10-deoxymethynolide production was abolished with low levels of narbonolide produced under 10-deoxymethynolide production conditions. This indicates that the 85K protein is generated from alternative translation of a *pikAIV* messenger RNA, leading to

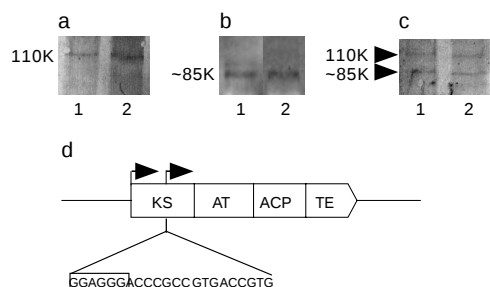


Figure 2 Alternative expression of *pikAIV*. **a–c**, Western blot analyses of *pikAIV* expression in *S. venezuelae* AX910 (Lane 1) and *S. venezuelae* AX912 (Lane 2). Narbonolide production conditions (**a**), 10-deoxymethynolide production conditions (**b**) and culture conditions corresponding to the production of both narbonolide and 10-deoxymethynolide (**c**). **d**, The alternative translational start sites for *pikAIV* expression. The expanded region is the downstream site with start codons (underlined) and the corresponding ribosome-binding region (in a box).

Genotype of the <i>pikAIV-pikAV</i> loci in recombinant <i>S. venezuelae</i> strains	
WT chromosome (control)	
AX912 chromosome (recombinant host)	
pDHS702	
pDHS704	
pDHS705	
pDHS706	
pDHS708	
pDHS707	

Figure 3 Plasmid complementation of *S. venezuelae* AX912 and polyketide production. The *pikA* genes (*pikAIV*–*pikAIII*) on chromosome not shown) are represented by open arrows with their start and stop positions (amino-acid numbers) indicated in parentheses. Asterisk and dot indicate the internal translation start site and the His₆ tag, respectively.

a truncated *PikAIV* that contains only half of the module 6 ketosynthase (KS₆) domain (Fig. 2d). Because the KS₆ domain is responsible for condensation of the final extender unit, the alternatively expressed PKS system, which is unable to catalyse this reaction, would only be capable of generating intermediate 2 (Fig. 1) and the 12-membered ring macrolactone 10-deoxymethynolide.

The N-terminal truncated form of *PikAIV* is unusual for a modular PKS. To probe its function, we expressed varying length forms of *pikAIV* and tested them for their capacity to complement AX912 (TE[−] TEII[−]). The results showed that the TE domain in *PikAIV* is critical for 10-deoxymethynolide formation (Fig. 3). Specifically, all of the constructs that contain the TE domain including pDHS704 (TE alone), pDHS705 (ACP₆-TE), pDHS706 (ACP₆-TE::TEII), pDHS708 (AT₆-ACP₆-TE) and pDHS707 (KS₆-AT₆-ACP₆-TE) complemented mutant AX912 to give 10-deoxymethynolide. However, the most efficient production of 10-deoxymethynolide resulted from complementation by pDHS708,

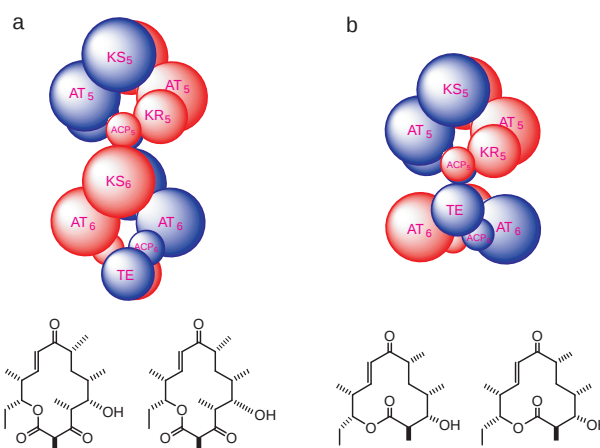


Figure 4 Structural models for *PikAIV* in 10-deoxymethynolide and narbonolide biosynthesis. **a**, Narbonolide production conditions; **b**, 10-deoxymethynolide production conditions. Spheres represent enzymatic domains in the PKSs with the diameter proportional to the size of the domain. Each PKS was dimerized (peptide chains are red and blue) and twisted 180° to form a half helix using the model for erythromycin PKS¹⁰, and then *PikAIII* domains were stacked on top of *PikAIV* according to their order in biosynthesis. Two sets of independent active sites are thus formed along two grooves of the helix that lead to the production of two polyketides in each biosynthetic cycle.

10-Deoxymethynolide production (SCM medium)	Narbonolide production (PGM medium)
100%	100%
0	0
trace	trace
~60%	0
~20%	trace
~30%	~10% narbonolide trace 10-deoxymethynolide
~100%	~50% 10-deoxymethynolide ~50% narbonolide
~40% 10-deoxymethynolide ~60% narbonolide	100%

Polyketide production in each recombinant system is the average production of 2–3 randomly picked primary transformants. The yields were normalized by using AX912 as 0% and full-length *pikAIV* complementation (pDHS707) as 100% production standards.

which contains AT₆-ACP₆-TE domains and closely mimics the truncated form of PikaIV detected in wild-type *S. venezuelae* under conditions for 10-deoxymethynolide production (Fig. 3). These results indicate that the TE domain may be the key active domain in the truncated PikaIV polypeptide, and the remaining domains in the N-terminal truncated form of PikaIV (including acyl-transferase (AT), acyl carrier protein (ACP), and probably a partial KS domain) are also important in effective production of polyketide products.

We developed a model on the basis of the proposed helical structure of the erythromycin PKS¹⁰ to illustrate the function of each domain in PikaIV during polyketide biosynthesis (Fig. 4). Under conditions for narbonolide production, wild-type *S. venezuelae* expresses a full-length PikaIV, which interacts with PikaIII and elongates the growing polyketide chain on ACP₅ by adding a methylmalonate unit (the activity of KS₆) to produce the 14-membered ring macrolactone (Fig. 4a). Under culture conditions for 10-deoxymethynolide production, the truncated form of PikaIV that lacks KS₆ is expressed. The molecular space left unoccupied by KS₆ truncation is then presumably filled by the TE domain that would be aligned to interact directly with ACP₅ to release the 12-membered ring macrolactone (Fig. 4b). In both cases, the main part of PikaIV is predicted to remain fixed. A small movement of the TE domain into the unoccupied space (left by KS₆ truncation) would result in bypass of the AT₆-ACP₆ catalytic domains in the truncated PikaIV, but retention of thioesterase activity. In this model, the remaining domains in PikaIV serve as a scaffold that orients the TE domain and stabilizes the interacting complex between PikaIII and PikaIV, thus greatly increasing the production of 10-deoxymethynolide as shown in the complementation experiments.

Our proposed mechanistic model is consistent with the results of the complementation study. The dimeric architecture of modular PKSs is apparent in the complementation by pDHS708 (AT₆-ACP₆-TE) in PGM medium, in which the system produces equal amounts of 10-deoxymethynolide and narbonolide. The production of narbonolide reflects the formation of a heterodimer between the KS₆-truncated form of PikaIV generated from the plasmid and a TE domain-deleted form of PikaIV from the mutant AX912 chromosome, whereas the production of 10-deoxymethynolide reflects a homodimer of the KS₆-truncated form of PikaIV from the complementation plasmid. The proposed movement of the TE domain in truncated PikaIV is derived from an inferred interaction between the TE domain and PikaIII (ACP₅). Complementation of *S. venezuelae* AX912 with pDHS704 (expressing TE domain alone) resulted in 60% production of 10-deoxymethynolide under culture conditions for 10-deoxymethynolide production, whereas the same complementation system did not result in macrolactone formation under culture conditions for narbonolide production. These results indicate that there may be a delicate balance in the molecular interactions between KS₆ and PikaIII (ACP₅) or between TE domain and PikaIII (ACP₅) with the KS₆-ACP₅ interaction being more dominant. When the PikaIV KS₆ domain is present under narbonolide culture conditions, this KS₆-ACP₅ interaction prevents the TE domain from interacting with ACP₅. Conversely, when the truncated form of PikaIV (KS₆) is formed under culture conditions for 10-deoxymethynolide production, the TE-PikaIII interaction prevails and 10-deoxymethynolide is produced. The result also suggests that the TE domain must be covalently linked to ACP₆ to release narbonolide, which is similar to an observation in the erythromycin system¹¹.

The complementation results involving the truncated forms of PikaIV provide new insights into the structure and function of these giant multi-enzyme systems. Efficient production of 10-deoxymethynolide by a truncated form of PikaIV indicates that the AT domain may have a pivotal role in PKS dimerization. The KS₆-truncated form of PikaIV expressed from pDHS708 (AT₆-

ACP₆-TE) efficiently generates hetero- and homodimers to produce 10-deoxymethynolide and narbonolide, but molecular interaction was severely limited when the AT₆ domain was truncated in pDHS705 (ACP₆-TE). This pivotal role is illustrated in the PKS structural model in which AT is at the centre of the protein dimer (Fig. 4). The KS domain (especially N-terminal part of the KS domain) also has a critical role in inter-PKS interaction (A similar conclusion has also been reached recently from the erythromycin system¹²). The dominant interaction between KS₆ and PikaIII (ACP₅) in the Pik system may serve as the primary basis of module-module recognition and docking in multifunctional PKS systems. The *pika* system in *S. venezuelae* provides a unique opportunity as well as a powerful tool to study these fundamental interactions in further detail.

The complementation results clarify further the function of PikaV (TEII). TEII alone is not sufficient for polyketide termination (as shown in pDHS702 complementation (Fig. 3) and in mutant AX910); however, the independent thioesterase did enhance the production of both 10-deoxymethynolide and narbonolide (Fig. 3). This accessory role of TEII is consistent with observations in other PKSs^{13,14} as well as in non-ribosomal peptide synthetase systems¹⁵.

It is valuable to compare alternative termination by differential expression of PikaIV in *S. venezuelae* with engineered polyketide chain-length manipulations from other PKS systems. In the erythromycin PKS, the TE domain from EryAIII was moved to upstream domains and covalently linked to alternative ACPs resulting in truncated polyketides^{16,17}. In each case, the capacity for producing the full-length polyketide product was subsequently eliminated. In contrast, by linking the TE domain of PikaIV to an upstream module by protein-protein interactions, *S. venezuelae* retains the capacity to generate two macrolactones of different size.

The differential expression of PikaIV is apparently regulated (directly or indirectly) by some factor(s) in the bacterial growth medium. In nature, this may correspond to ecological conditions in which different antibiotics are necessary. Therefore, controlling macrolactone structure by alternative gene expression may be a mechanism by which *S. venezuelae* administers the effective use of its antibiotics. Moreover, this regulatory scheme and the capacity to generate multiple macrolactones in the *pika* system may be readily harnessed for combinatorial biosynthesis of new natural products. □

Methods

Genetic manipulation, antibiotic extraction and western blots

Mutant AX910 and AX912 were created by targeted gene replacement. The designed mutations were first created on plasmids (pDHS910 and pDHS912, respectively) and then introduced into *S. venezuelae* chromosome by homologous recombinations between the plasmids and bacterial chromosome as described⁹. Polyketide extraction and quantification followed established procedures⁷. Western blot analyses of PikaIV expression in *S. venezuelae* AX910 and AX912 followed standard protocols¹⁸ and were conducted using anti-His₆ antibody (Qiagen) against the total protein extracts from 3–4-day cultures in the corresponding medium.

Construction of complementation plasmids

We designed a low-copy number expression plasmid (based on SCP2* origin of replication¹⁹) with the target gene under the control of the *pikaI* promoter⁹ so that expression of *pikaIV* variants from the plasmid would most closely resemble a normal temporal expression profile, and would also be synchronized with expression of the *pika* cluster encoded on the *S. venezuelae* chromosome. The *pikaI* promoter, *PpikaI*, was isolated as an EcoRV–EcoRI fragment between *pikaI* and *pikaR1* in the *pika* cluster⁹. To create a plasmid for complementation, a DNA fragment including *pikaIV* was initially amplified by polymerase chain reaction (PCR) and placed downstream of the EcoRI site so that *pikaIV* was translationally coupled to the leader sequence of *pikaI* in *PpikaI* to give plasmid pDHS702. Plasmids pDHS904, pDHS705, pDHS706, pDHS707 and pDHS708 were constructed by cloning various lengths of the *pikaIV*–*pikaIV* region into pDHS702 and replacing *pikaIV*.

Site-directed mutagenesis of *pikaIV*

Mutation of potential alternative start codons (GTGs) in *pikaIV* was conducted using Change-A-Base system (Stratagene). The PCR primer pair used in this process is 5'–

TGGAGGGACCCGCCGTCACCGTCGACACGGCGTGCTC-3' and 5'-GAGCAGCC GTGTCGACGGTGACGGCGGGTCCCTCCA-3'. The mutations were confirmed by sequencing before and after introduction into *S. venezuelae*.

Received 27 September; accepted 22 November 1999.

1. Hopwood, D. A. & Sherman, D. H. Molecular genetics of polyketides and its comparison to fatty acid biosynthesis. *Annu. Rev. Genet.* **24**, 37–66 (1990).
2. Cane, D. E., Walsh, C. T. & Khosla, C. Harnessing the biosynthetic code: combinations, permutations, and mutations. *Science* **282**, 63–68 (1998).
3. Donadio, S., Staver, M. J., McAlpine, J. B., Swanson, S. J. & Katz, L. Modular organization of genes required for complex polyketide biosynthesis. *Science* **252**, 675–679 (1991).
4. Cortes, J., Haydock, S. E., Roberts, G. A., Bevtit, D. J. & Leadlay, P. F. An unusually large multifunctional polypeptide in the erythromycin-producing polyketide synthase of *Saccharopolyspora erythraea*. *Nature* **348**, 176–178 (1990).
5. Lambalot, R. H. & Cane, D. E. Isolation and characterization of 10-deoxymethynolide produced by *Streptomyces venezuelae*. *J. Antibiot.* **45**, 1981–1982 (1992).
6. Hori, T., Maezawa, I., Nagahama, N. & Suzuki, M. Isolation and structure of narbonolide, narbomycin aglycone, from *Streptomyces venezuelae* and its biological transformation into picromycin via narbomycin. *J. Chem. Soc. Chem. Commun.* 304–305 (1971).
7. Sherman, D. H. *et al.* in *Proceedings of the 8th International Biotechnology Symposium* Vol. 1 123–137 (Societe Francaise de Microbiologie, Paris, 1988).
8. Tang, L., Fu, H., Betlach, M. C. & McDaniel, R. Elucidating the mechanism of chain termination switching in the picromycin/methymycin polyketide synthase. *Chem. Biol.* **6**, 553–558 (1999).
9. Xue, Y., Zhao, L., H. W. & Sherman, D. H. A gene cluster for macrolide antibiotic biosynthesis in *Streptomyces venezuelae*: architecture of metabolic diversity. *Proc. Natl Acad. Sci. USA* **95**, 12111–12116 (1998).
10. Staunton, J. *et al.* Evidence for a double-helical structure for modular polyketide synthases. *Nature Struct. Biol.* **3**, 188–192 (1996).
11. Gokhale, R. S., Hunziker, D., Cane, D. E. & Khosla, C. Mechanism and specificity of the terminal thioesterase domain from the erythromycin polyketide synthase. *Chem. Biol.* **6**, 117–125 (1999).
12. Gokhale, R. S., Tsuji, S. Y., Cane, D. E. & Khosla, C. Dissecting and exploiting intermodular communication in polyketide synthases. *Science* **284**, 482–485 (1999).
13. Rangaswamy, V., Mitchell, R., Ullrich, M. & Bender, C. Analysis of genes involved in biosynthesis of coronafacic acid, the polyketide component of the phytotoxin coronatine. *J. Bacteriol.* **180**, 3330–3338 (1998).
14. Butler, A. R., Bate, N. & Cundliffe, E. Impact of thioesterase activity on tylosin biosynthesis in *Streptomyces fradiae*. *Chem. Biol.* **6**, 287–292 (1999).
15. Schneider, A. & Marahiel, M. A. Genetic evidence for a role of thioesterase domains, integrated in or associated with peptide synthetases, in non-ribosomal peptide biosynthesis in *Bacillus subtilis*. *Arch. Microbiol.* **169**, 404–410 (1998).
16. Cortes, J. *et al.* Repositioning of a domain in a modular polyketide synthase to promote specific chain cleavage. *Science* **268**, 1487–1489 (1995).

17. Kao, C. M., Luo, G. L., Katz, L., Cane, D. E. & Khosla, C. Manipulation of macrolide ring size by directed mutagenesis of a modular polyketide synthase. *J. Am. Chem. Soc.* **117**, 9105–9106 (1995).
18. Sambrook, J., Fritsch, E. F. & Maniatis, T. *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1989).
19. Lydiat, D. J., Malpartida, F. & Hopwood, D. A. The *Streptomyces* plasmid SCP2⁺: its functional analysis and development into useful cloning vectors. *Gene* **35**, 223–235 (1985).
20. Bisang, C. *et al.* A chain initiation factor common to both modular and aromatic polyketide synthases. *Nature* **401**, 502–505 (1999).
21. Xue, Y., Wilson, D., Zhao, L., Liu, H.-w. & Sherman, D. H. Hydroxylation of macrolactones YC-17 and narbomycin is mediated by the *pikC*-encoded cytochrome P450 in *Streptomyces venezuelae*. *Chem. Biol.* **5**, 661–667 (1998).

Acknowledgements

We thank H.-w. Liu, K. A. Reynolds and G. M. Dunny for comments on the manuscript, and X. He for expert technical assistance. This research was supported by the NIH and a grant from the Office of Naval Research (to D.H.S.).

Correspondence and requests for materials should be addressed to D.H.S. (e-mail: david-s@biosci.umn.edu).

erratum

The neurobiology of cognition

M. James Nichols & William T. Newsome

Nature **402**(Suppl.), C35–C38 (1999)

Figure 3a of this paper was published without proper attribution. The photograph was taken by the late F. W. Goro, and copyright belongs to his estate. The editors apologize to the executors of the estate for this oversight.

Six chromatography columns

Recent releases from manufacturers of chromatographs and accessories.

761 Chromatography System

Metrohm www.metrohm.ch

A compact approach

The Metrohm 761 Compact IC has a footprint of only one-and-a-half A4 sheets and a height of only 45 cm, but is designed to cope with all the demands of ion chromatography. There is a choice of three different measuring ranges for optimal signal-to-noise ratio. All functions can be controlled from a PC, but the PC does not require additional hardware: 20 MB free memory, Pentium or higher, and 64 MB internal memory will suffice.

Reader Service No. 100



Low cunning: 45 cm of Metrohm chromatography.

but fully backward-compatible to NovaPrep systems running under Windows 95.

Reader Service No. 105

Parallex Flex

Biotage www.biotage.com

HPLC for combinatorial chemistry

The new Parallex Flex system from Biotage, a division of Dyax Corp, is designed for flexibility in HPLC purification in the combinatorial chemistry lab. The fully automated system runs from one to four HPLC channels in parallel, each with its own pump to ensure uniform flow. The sample injector is capable of running 768 samples using up to eight deep-well microtitre plates or racks of tubes, and each column has its own fraction collector.

Reader Service No. 106

Parallel LC-MS

Micromass www.micromass.co.uk

Pieces of eight

Micromass claim that this system, launched at PittCon '99, was the first automated parallel LC-MS system on the market. Up to 8 parallel HPLC channels are possible with a MUX interface allowing 8 sample streams to be multiplexed (MUX) with one Time-Of-Flight MS detector. The MassLynx NT data system tracks the data from each liquid stream separately in a 32-bit Windows NT environment.

Reader Service No. 101

µ-Flow

SGE Europe www.sge.com

Waste not, want not

µ-Flow provides accurate and reproducible microflows of 1–250 µl per min and enables a conventional HPLC isocratic piston pump to deliver micro volume flows suitable for 0.3–2-mm ID columns. With this splitter you can achieve the same results as those obtained using a specialized micro HPLC pump. Waste solvent from the splitter can be recycled, saving solvent and money.

Reader Service No. 103

Training programs

Savant

A bilingual (Spanish/English) computer-based training program

Using *Introducción a la Cromatografía de Gases*, users can change instantly from English to Spanish or Spanish to English by just clicking on a menu item. In addition to being a tool for teaching GC, the program offers an easy way to learn technical GC terms in English or Spanish. The self-paced program features a chromatographic simulator which shows results of changes on the chromatogram when experimental parameters are varied.

Reader Service No. 107

UniJet columns

BAS www.bioanalytical.com

For fast LC and other applications requiring metal-free conditions

These aluminium-encased PEEK reversed-phase narrow-bore columns have similar characteristics to the established BAS Phase II cartridge column, but give a longer life and higher plate count — over 120,000 plates per metre. UniJet columns are packed with a 3-µm spherical ODS silica, and come in a 100-mm length and 3-mm or 2-mm ID formats. A fast LC version, with 50-mm length, 2-mm ID and 5-µm bonded silica, intended for use with LC/MS, is also part of the range.

Reader Service No. 102



PEEK performance: metals outside, not inside.

Dispensette

BrandTech Scientific www.brandtech.com

Dispensing aid for HPLC

The Dispensette Organic uses the same 'floating piston' mechanism as the Dispensette III bottle-top dispenser, but is made from materials that give better long-term compatibility with organic solvents, and with concentrated acids. The instrument is not suited to alkaline solutions or HF. The Dispensette Organic is recommended for HPLC labs, as it is suitable for TFA, THF, methylene chloride, acetonitrile and methanol used in many mobile phases.

Reader Service No. 104

LC R&Sponder

Hitachi Instruments, Inc. www.hii-hitachi.com

Preparative chromatography software for NovaPrep HPLC systems

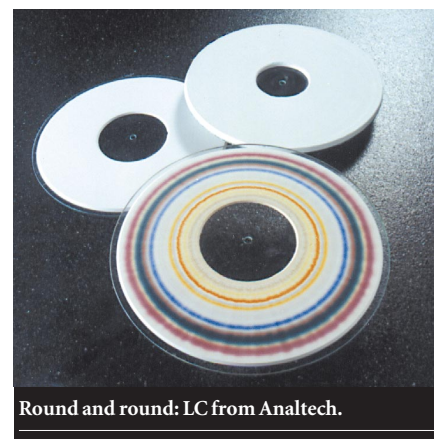
This software is optimized for preparative chromatography, and employs the same software that is used on the company's process scale systems. This allows an easy transfer of methods from the prep system to scale-up. LC R&Sponder is Windows NT 4.0 based,

TLC rotors

Analtech www.thinlayer.com

High-performance glass rotors

These new pre-cast silica gel rotors are designed for use in Chromatotron and Cyclo-Graph centrifugal chromatography systems.



Round and round: LC from Analtech.

new on the market



Flash equipment: FlashMaster from Jones.

Made of precision-cut glass, these high-performance rotors allow minimum diffusion of separated bands while maximizing flow rates. The rotors, along with scrapers, are available in 1,000, 2,000, 4,000, 6,000 and 8,000 μm nominal thickness. A 254-nm UV fluorescent indicator is included in all rotor coatings to facilitate easy viewing of the separations.

Reader Service No. 108

OX-700

Zinsser

www.zinsser-analytic.com

Automated quench-free sample oxidation for liquid/solid chromatography

For the preparation of ^3H - or ^{14}C -labelled biological samples, combustion is still the method of choice. The OX-700 catalytic Sample Oxidizer can handle a wide range of materials in this technique. At over 900 $^{\circ}\text{C}$ it converts the sample into $^3\text{H}_2\text{O}$ and $^{14}\text{CO}_2$ and traps it in separate scintillation vials. The catalytic oxidation ensures removal of quenching agents such as SO_2 and NO_2 .

Reader Service No. 109

FlashMaster

Jones Chromatography

www.jones-chrom.co.uk

Automated chromatography

The FlashMaster PC-based system offers multiple sample operation and intelligent fraction collection with sample tracking.

FlashMaster features disposable pre-packed columns that provide easy sample loading and eliminates the need to wash and re-pack glass columns. Gradient pumping capability enables the FlashMaster to achieve reproducible flow rates and reduce costly solvent consumption.

Reader Service No. 110

TSKgel MultiporeH_{XL}-M column

TosoHaas

www.tosohaas.com

A column for gel-permeation chromatography

The TSKgel MultiporeH_{XL}-M column offers a novel packing material for separating polymers by GPC. The PS/DVB resin contains pores of multiple size continuously distributed throughout the resin. The multi-pore resin has advantages over mixed-bed columns containing blends of different-grade resins. They can attain better linearity on calibration curves over a wide MW range, and achieve smooth chromatograms without inflections often seen with mixed-bed columns.

Reader Service No. 111

Reporting tool

Blaise Software

www.blaisesoft.com

A reporting system for advanced data handling

A deal between PerkinElmer Instruments and Blaise Software Services, Inc. means that Blaise's WinBSR 2.0 reporting system is now incorporated into the Turbochrom Workstation and Turbochrom Client/Server Chromatography Data Systems. The reporting system, known as TC Publisher, includes Microsoft Visual Basic for Applications, providing custom calculations and custom report formatting, including WYSIWYG report preview.

Reader Service No. 112

HPLC

Polymer Laboratories www.polymerlabs.com

Narrow-bore HPLC column hardware

This new style of 2.1-mm ID stainless-steel column hardware has been designed to



Narrow-minded: small-bore HPLC.

minimize dead volume in the end fitting and therefore improve efficiency and symmetry. The reduction in internal diameter from the conventional 4.6-mm ID columns to narrow-bore 2.1-mm ID hardware provides a reduction in band broadening, as reduced volumetric flow rates can be used. This is particularly advantageous for LC/MS analyses and identification of trace analytes, as an increase in signal-to-noise ratio is obtained.

Reader Service No. 113

Analytical nitrogen

Whatman

www.whatman.com

Nitrogen on tap for LC/MS

The Analytical Nitrogen Gas Generator from Whatman is designed to supply a continuous and consistent flow of high-quality dry nitrogen to LC/MS systems and other analytical instruments. The N2-2010 model uses membrane technology to separate N_2 from compressed air, delivering gas at a purity of 99.9%. Several instruments can receive gas from the unit simultaneously, depending on the total gas-flow requirement. The generator removes the need to change cylinders, and the threat of gas running out in the middle of a run.

Reader Service No. 114

ms-NoVent

SGE

Change chromatography gear

The ms-NoVent is an interface for mass spectrometers designed to allow column changing or maintenance without the need to shut down the MS. The ms-NoVent, installed directly onto the inlet of the ion source, will prevent air entering the mass spectrometer while the column is disconnected. So the ion source remains contamination-free and the mass spectrometer is operational almost immediately.

Reader Service No. 115

These notes are compiled in the Nature office from information provided by the manufacturers. For more details, fill in the reader service card.

ADVERTISEMENTS

Check out our homepage at <http://www.resgen.com>

**BACs
YACs
PACs
cDNAs**

Research Genetics provides clones and colony membranes for a number of human, mouse, and rat libraries as well as a custom screening service for the libraries.

Research Genetics, Inc.
U.S. or Canada 800-533-4363
U.K. 0-800-89-1393
FAX 256-536-9016

READER ENQUIRY NO. 6

RT activity

the best assays

www.cavidi.com

CAVIDI TECH

READER ENQUIRY NO. 29